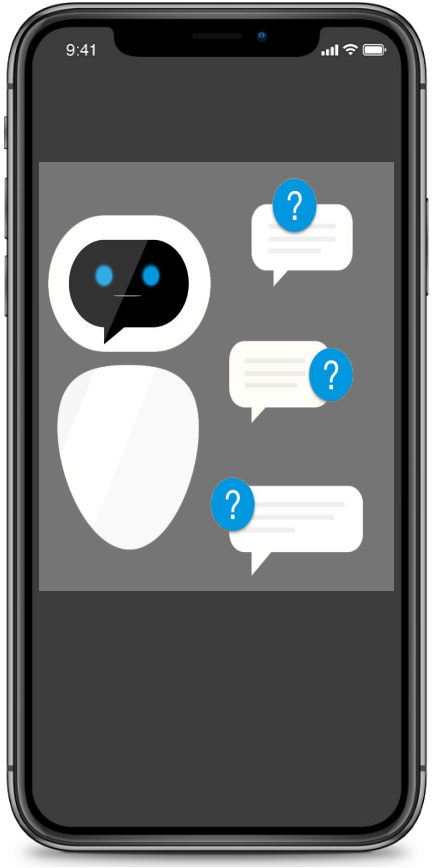


QBE DATA SCIENCE WORKSHOP

Data Analytics

Review on Data Unification Activity

Quintegral

General Framework

Solving the Problem

- identify the task -- why?
- scope if this is possible
  - do you have the data?
  - is it possible with the tools I have?
  - what techniques can you use to solve this problem?
- lay the general steps to implement the solution
- (if necessary) gather data
- code
- test

QBE

Data Unification

Solving the Problem

- identify the task: **unify the dataset for proper assessment of the status of the tickets being raised by customers**
- scope if this is possible
  - do you have the data? (**name, email, contact number, orderID, ticketID)**
  - is it possible with the tools I have? **Python, Pandas**
  - what techniques can you use to solve this problem? **Sets, Networks**

Input: **Customer_Tickets(name, email, contact number, orderID, ticketID)**
Output: **Cleaned dataset**

	Id	Email	Phone	Contacts	OrderId	Status	CSR	Datetime	Location
0	0	gkzAbly@qq.com	NaN	1	NaN	Pending	JPM	02/05/2021 13:50	Las Pinas
1	1	NaN	3.294430e+11	4	vDDJJcxfLtSfkooPhbYnJdxov	Open	K02	29/10/2020 8:46	Pasig
2	2	NaN	9.125984e+09	0	NaN	Resolved	MEO	29/12/2020 4:10	Quezon City
3	3	mdllpYmE@gmail.com	NaN	0	bHquEnCbbsGLqllwryxPsNOxa	Closed	I9S	26/10/2020 2:04	Taguig
4	4	NaN	3.003644e+08	2	NaN	Resolved	CA3	04/03/2021 11:47	Valenzuela

Data Unification

Solving the Problem

- lay the general steps to implement the solution
  - **load the dataset**
  - **data engg: clean the dataset**
  - **unify the dataset**
    - **unify same email -> A**
    - **unify same contact number -> B**
    - **unify same order ID -> C**
    - **unify A, B, C**
- (if necessary) gather data : **available**
- code
- test
- analyze

* need to set a new ID since the previous IDs still have duplicates
* need another table to fill the new/combined dataset

QBE

Data Unification

Solving the Problem

code

- **load the dataset**

```
df = pd.read_csv('./csr_operations_data.csv')
```

- **data engg: clean the dataset**

```
# replace all ' ' strings with NaN value
df = df.replace(r'^\s*$', np.NaN, regex=True)
```

- **unify the dataset**
  - unify same email -> A
  - unify same contact number -> B
  - unify same order ID -> C
  - *need to set new IDs
  - (cont.)

```
email_group = df.groupby('Email').Id.agg(lambda x: set(x))
phone_group = df.groupby('Phone').Id.agg(lambda x: set(x))
order_group = df.groupby('OrderId').Id.agg(lambda x: set(x))
```

```
for ids in email_group:
    for id in ids:
        d[id] |= set(ids)
for ids in phone_group:
    for id in ids:
        d[id] |= set(ids)
for ids in order_group:
    for id in ids:
        d[id] |= set(ids)
```

QBE

Data Unification

Solving the Problem

- **unify the dataset**
  - (cont.)
  - unify A, B, C

```
for i in tqdm(range(3)):
    for id, ids in d.items():
        for id_ in list(ids):
            d[id] |= d[id_]
```

  - save new IDs

```
id_to_contact = df.set_index('Id').Contacts.to_dict()
```

  - define a function that will get the sum of all the contacts for the new set of users (unified)

```
def get_sum_contact(ids_set):
    return sum([id_to_contact[id] for id in ids_set])
```

  - create another variable to store the combined dataset

```
df['set'] = df.Id.apply(lambda x: d[x])
df['trace'] = df.set.apply(lambda x: '-'.join(map(str, sorted(list(x)))))
df['n_con'] = df.set.apply(lambda x: str(get_sum_contact(x)))
df['out'] = df.trace + ', ' + df.n_con
out = df[['Id', 'out']]
out.columns = ['ticket_id', 'ticket_trace/contact']
```

  - export/save data

```
out.to_csv('out.csv', index=False)
```

Data Unification

(4) CODE

* can code per bullet: A, B, C.1, C.2, ...
* by breaking down the steps, you can find fxns/libraries or codes that can do them
* GOOGLE IS YOUR FRIEND!

(5) TEST

* check if the outputs are correct
* you can use a small subset of data for checking

QBE

Data Unification

Solving the Problem

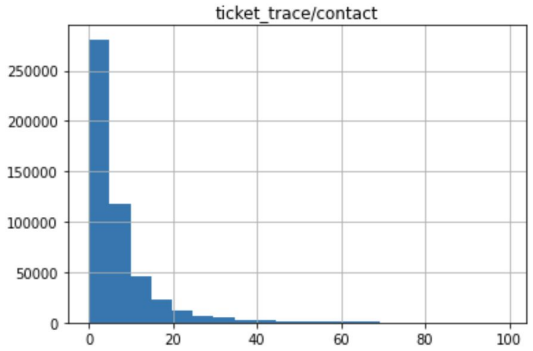
- analyze

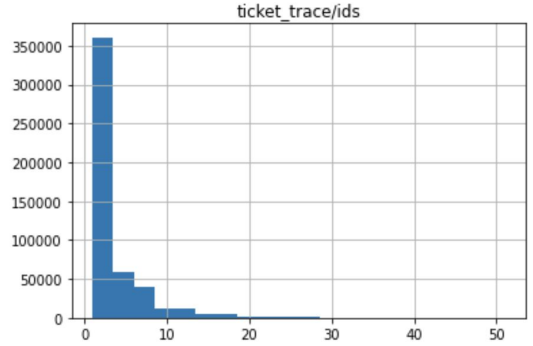
```
out_analysis = pd.DataFrame()
out_analysis['ticket_id'] = df.trace
out_analysis['ticket_trace/contact'] = df.n_con

out_analysis['ticket_trace/ids'] = out_analysis.ticket_id.str.count("-")+1
#out_analysis
```

```
out_analysis.head(5)
```

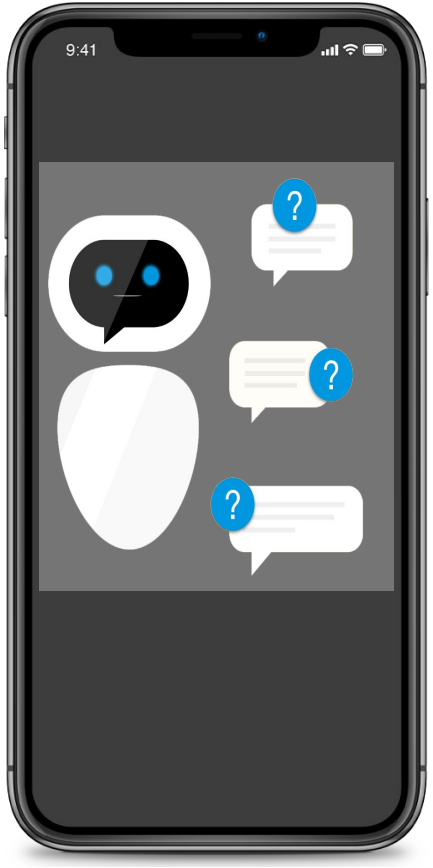
	ticket_id	ticket_trace/contact	ticket_trace/ids
0	0	1	1
1	1-2458-98519-115061-140081-165605-476346	12	7
2	2-159312-322639-348955	4	4
3	3	0	1
4	4	2	1





General Framework

- identify the task
- scope
- design the solution
- code
- test
- *deploy*



QBE DATA SCIENCE WORKSHOP

Introduction to Natural Language Processing (NLP)

How NLP is leveraged in expanding business

Quintegral

Table of Contents

POINTS FOR DISCUSSION:

- What is Natural Language Processing (NLP)?
- Framework and infrastructures
- Challenges in Computer Vision
- Applications
- Is remote work here to stay?
- What's next?

Quintegral

Communication

Typical communication episode

S (speaker) wants to convey P (proposition) to H (hearer) using W (words in a formal or natural language)

Speaker

- **Intention**: S wants H to believe P
- **Generation**: S chooses words W
- **Synthesis**: S utters words W

Hearer

- **Perception**: H perceives words W” (ideally W” = W)
- **Analysis**: H infers possible meanings P1,P2,…,Pn for W”
- **Disambiguation**: H infers that S intended to convey Pi (ideally Pi=P)
- **Incorporation:** H decides to believe or disbelieve Pi

Natural Language Processing

MonkeyLearn

Very intuitive platform, I'll definitely recommend it. The chat support is excellent, really fast in their replies and very helpful.

Usability | Positive | Customer Support

understanding

Taking some spoken/typed sentence and working out what it means

Quintegral

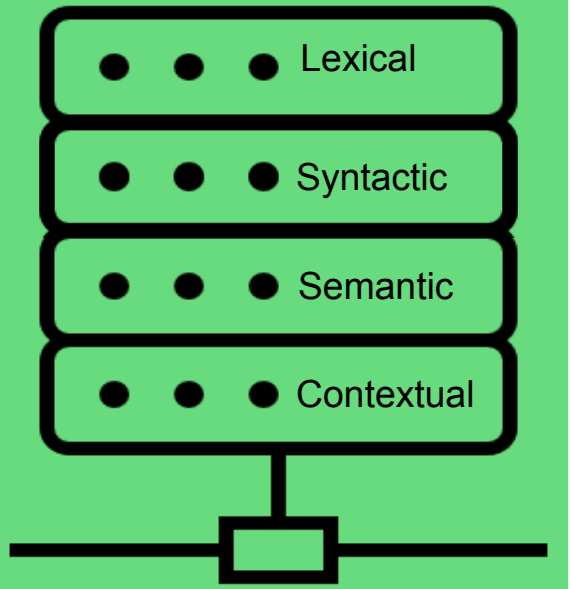
Natural Language Processing

generation

Taking some formal representation of what you want to say and working out a way to express it in a natural (human) language (e.g., English)

Quintegral

NLP Layers



Basic properties of words	Spell check, NER
Order and structure of words	Grammar check
Meaning of words	WordNet, etc.
Overall meaning of text	Topic modeling, sentiment analysis

Natural Language Understanding

01 Raw Speech Signal

Speech Recognition

↓

02 Sequence of words spoken

Syntactic analysis *using knowledge of the grammar*

↓

03 Structure of the sentence

Semantic analysis *using the info about the meaning of words*

↓

04 Partial representation of the meaning of sentence

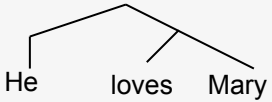
Pragmatic analysis *using info about context*

↓

05 Final Representation of meaning of sentence

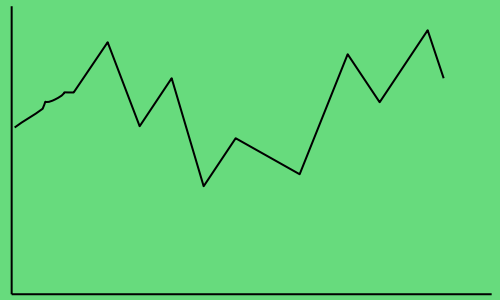
Quintegral

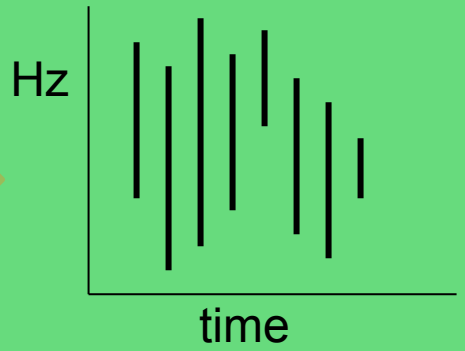
Natural Language Understanding

Input/Output data		Processing Stage	Other data used
Frequency spectrogram		speech recognition	Frequency of different sounds
Word sequence	“He loves Mary”	syntactic analysis	Grammar of language
Sentence structure		semantic analysis	Meanings of words
Partial Meaning	Ex loves(x,mary)	pragmatics	Context of utterance
Sentence meaning	loves(john,mary)		

Quintegral

Speech Recognition





Input
Microphone records voice

Analog Signal

Frequency spectrogram
e.g. Fourier transform

Quintegral

Speech Recognition

Typical communication episode

- Frequency spectrogram (basic sound signals, e.g. phonemes)
- Words

Complications

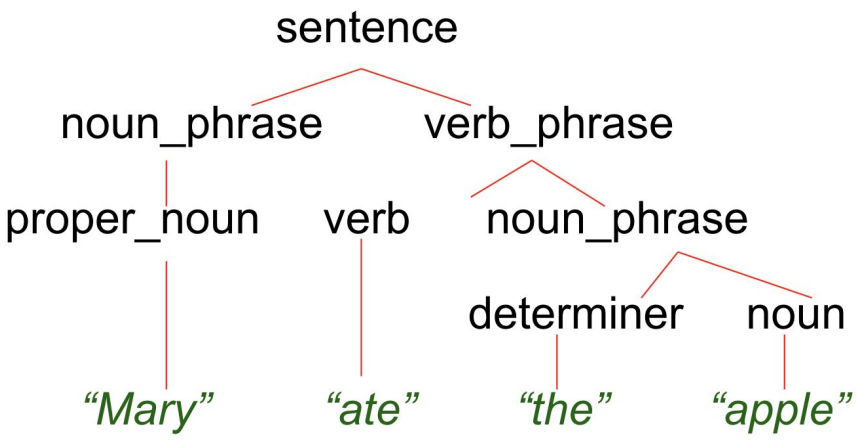
- No simple mapping between sounds and words
  - Variance in pronunciation due to gender, dialect, …
    - Restriction to handle just one speaker
  - Same sound corresponding to diff. words
    - e.g. bear, bare
  - Finding gaps between words
    - “how to recognize speech”
    - “how to wreck a nice beach”
  - Noise

Quintegral

Syntactic Analysis

Complications

- Rules of syntax (grammar) specify the possible organization of words in sentences and allows us to determine sentence's structure(s)
  - "I saw Mary with a telescope"
    - I saw (the man with a telescope)
    - I (saw the man with a telescope)
- Parsing: given a sentence and a grammar
  - Checks that the sentence is correct according with the grammar and if so returns a parse tree representing the structure of the sentence



Quintegral

Syntactic Analysis

Complications

- Syntatics ambiguity
  - `"Fruit flies like a banana."`
- Gerunds and adjectives
  - `"Frightening kids can cause trouble."`
- Having to parse syntactically incorrect sentences
  - `"John talked drugs to the children about."`

Semantic Analysis

Complications

- Handling ambiguity
  - Semantic ambiguity: “I saw the prudential building flying into Boston”

Newspaper Headlines

- Ban on Nude Dancing on Governor's Desk
- Iraqi Head Seeks Arms
- Juvenile Court to Try Shooting Defendant
- Teacher Strikes Idle Kids
- Stolen Painting Found by Tree
- Local High School Dropouts Cut in Half
- Red Tape Holds Up New Bridges
- Clinton Wins on Budget, but More Lies Ahead
- Hospitals Are Sued by 7 Foot Doctors
- Kids Make Nutritious Snacks

Pragmatics

Complications

- Uses context of utterance
  - Where, by who, to whom, why, when it was said
  - Intentions: inform, request, promise, criticize, …
- Handling Pronouns
  - “Mary eats apples. She likes them.”
    - She=“Mary”, them=“apples”.
- Handling ambiguity
  - Pragmatic ambiguity: “you’re late”: What’s the speaker’s intention: informing or criticizing?

Natural Language Processing

generation

Taking some formal representation of what you want to say and working out a way to express it in a natural (human) language (e.g., English)

Quintegral

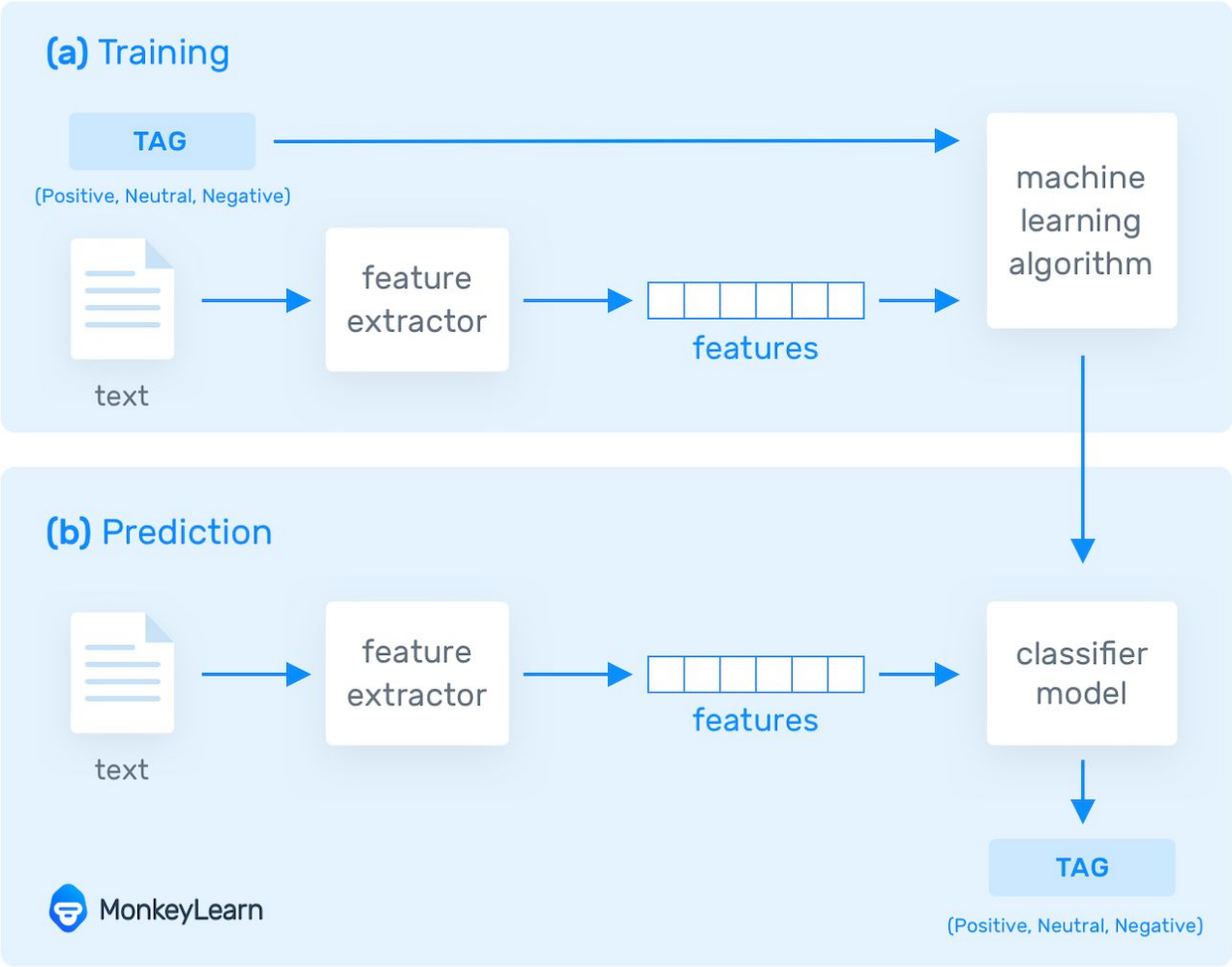
Natural Language Generation

- Talking back! ☺
- What to say or text planning
  - flight(AA,london,boston,$560,2pm),
  - flight(BA,london,boston,$640,10am),
- How to say it
  - “There are two flights from London to Boston. The first one is with American Airlines, leaves at 2 pm, and costs $560 …”
- Speech synthesis
  - Simple: Human recordings of basic templates
  - More complex: string together phonemes in phonetic spelling of each word
    - Difficult due to stress, intonation, timing, liaisons between words

Quintegral

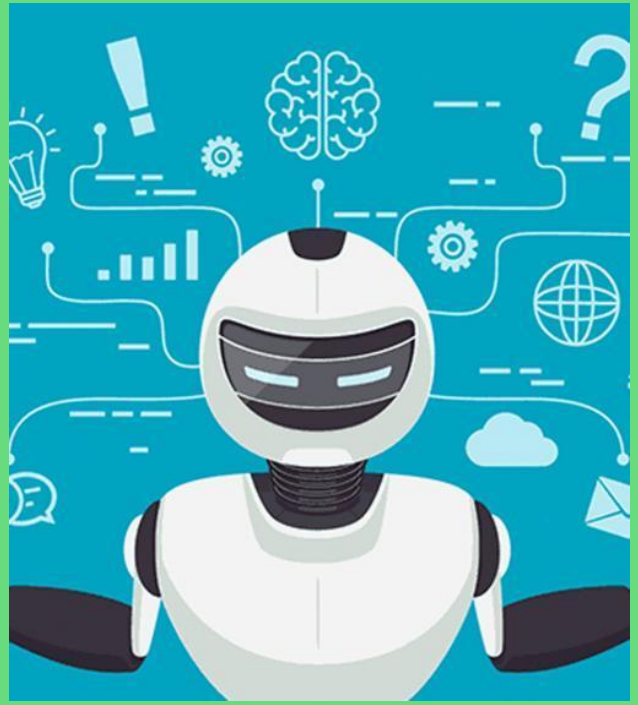
Natural Language Understanding

task: classifying



Universal question being addressed by NLP

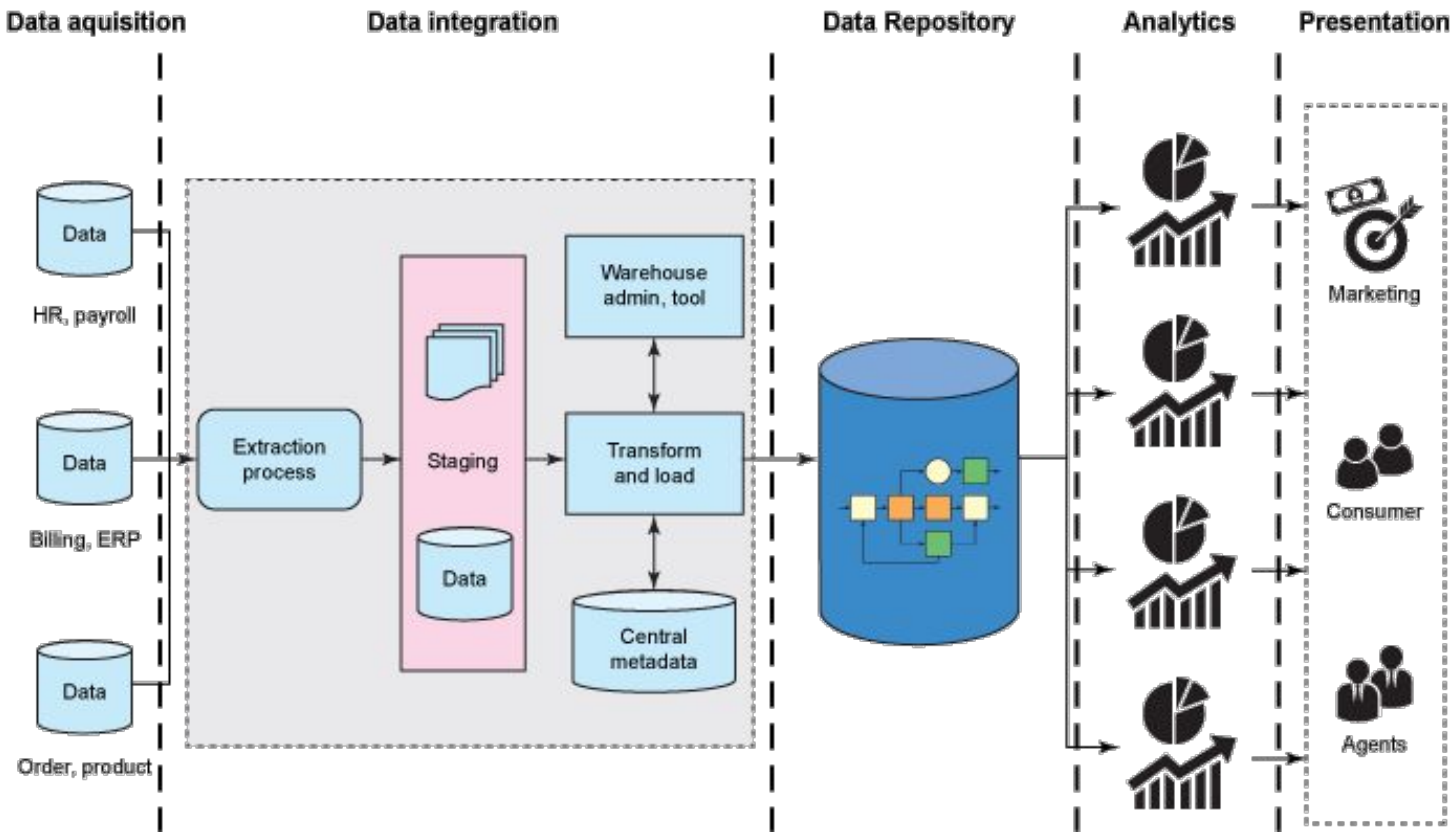
What does it really mean?



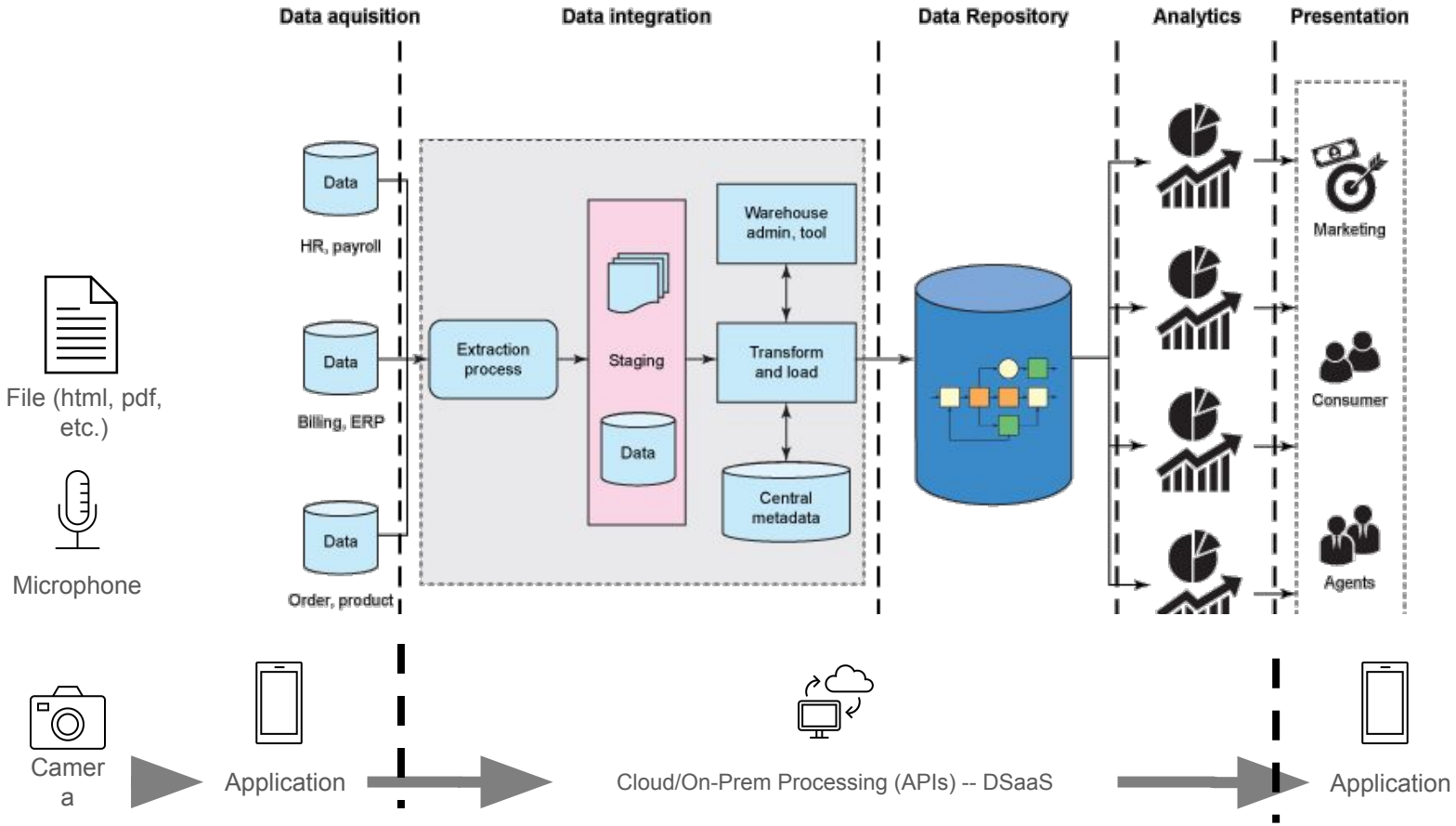
Components of a standard NLP system

- Data Source / Reference
- Context
- Processing Software
- Application (display screen, user-interface)

General Framework



General Framework

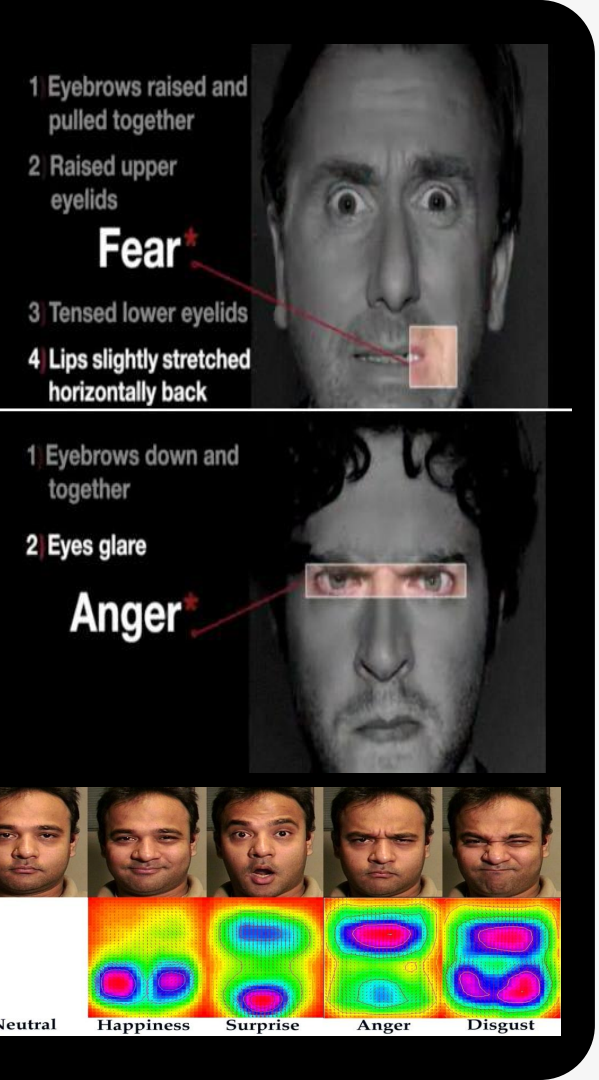


Proof of Concept (PoC)

edureka!

https://www.edureka.co/blog/what-is-machine-learnin

Processing Software



Pattern Recognition

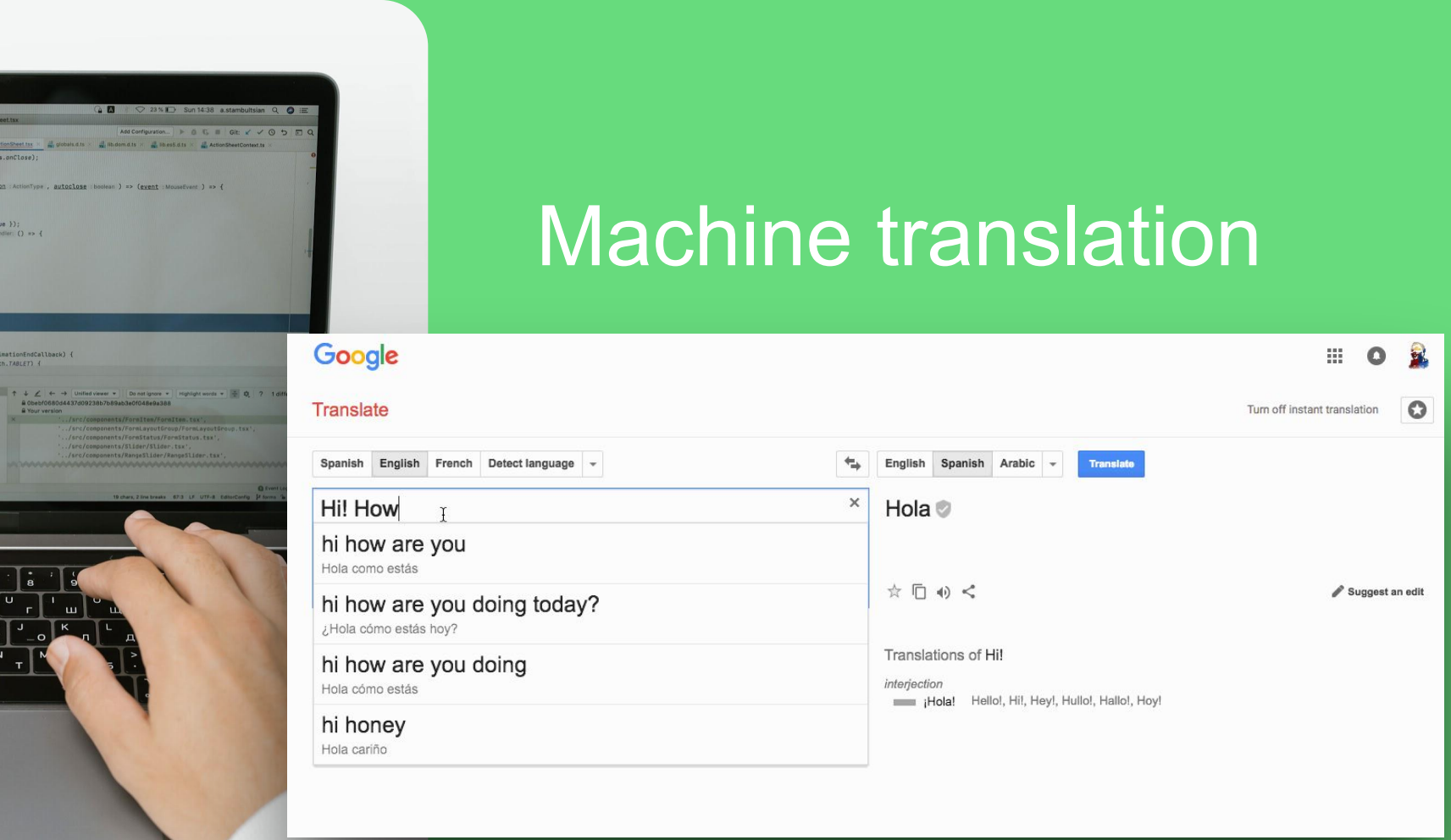
Machine learning Algorithms for Pattern Recognition

Avoid compromising the security of the platforms you use within the company by investing heavily in cybersecurity. Secure your company's data by looking into the best platforms out there that can protect against cybercriminals, hackers, and other fraudulent acts.

Applications

- Machine Translation
- Information Retrieval
  - Selecting from a set of documents the ones that are relevant to a query
- **Text Categorization**
  - Sorting text into fixed topic categories
- **Extracting data from text**
  - Converting unstructured text into structure data
- Spelling and grammar checkers
- Text summarization
- **Sentiment Analysis**

Machine translation



Information extraction

IE

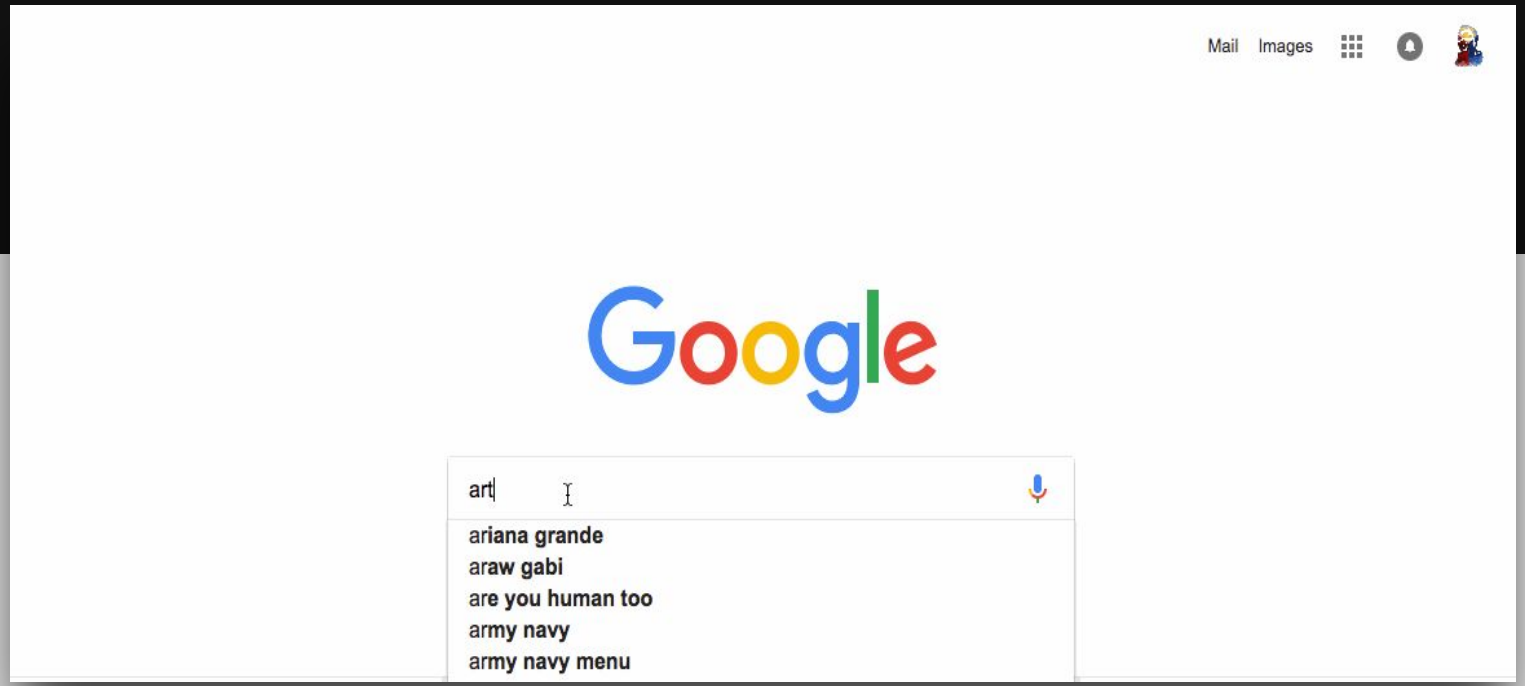
Subject	Relation	Object
p53	is_a	protein
Bax	is_a	protein
p53	has_function	apoptosis
Bax	has_function	induction
apoptosis	involved_in	cell_death
Bax	is_in	mitochondrial outer membrane
Bax	is_in	cytoplasm
apoptosis	related_to	caspase activation
...	...	...

textual abstract:
summary for human

structured knowledge extraction:
summary for machine

https://searchcio.techtarget.com/definition/Digital-enterprise

Google Search



https://indatalabs.com/blog/data-science/applications-of-natural-language-processing-in-business

Sentiment analysis

LEXALYTICS Platform Industries Applications Tech Demo AI Labs Blog Contact Support Login

immediately have a better idea of why consumers aren't happy.

This is the kind of insight one aims to find through market research, but why devote enormous budgets and countless man-hours to conducting surveys and cold calling? Through Lexalytics text mining tools, you'll get answers in seconds from the comfort of your office chair.

Below you will find a high-level overview of intentions and Lexalytics' sentiment analysis tools. For a more in-depth explanation of our sentiment analysis, read through "How Lexalytics Does Sentiment Analysis". Check out our web demo to see Lexalytics in action, or get in touch to schedule a live demo.

Get a FREE trial | Schedule full demo

Multi-level Analysis

Lexalytics' sentiment analysis tools can be configured to determine sentiment on a range of levels. We'll score sentiment on a document level (does this express a general positive or negative tone), but we'll also score the sentiment of individual words or phrases in the document.

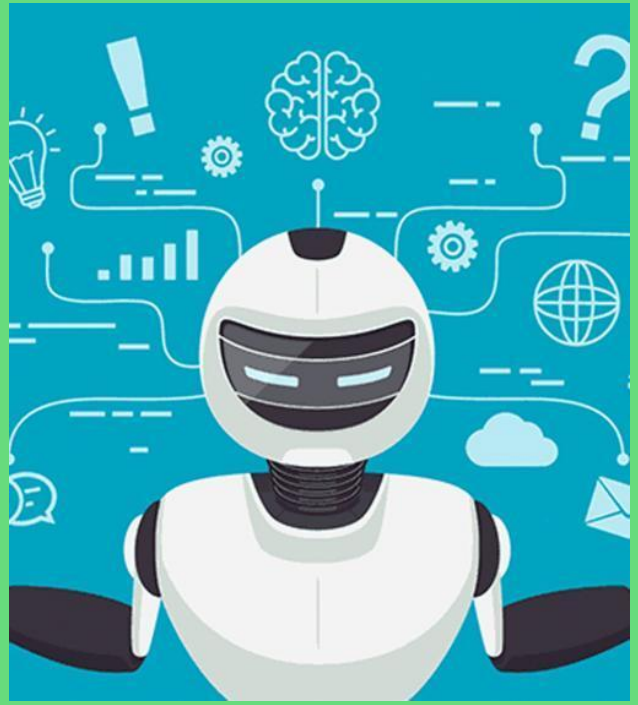
The process of determining whether a piece of writing is positive, negative or neutral.

It's also known as opinion mining

https://indatalabs.com/blog/data-science/applications-of-natural-language-processing-in-business

Why is NLP hard?

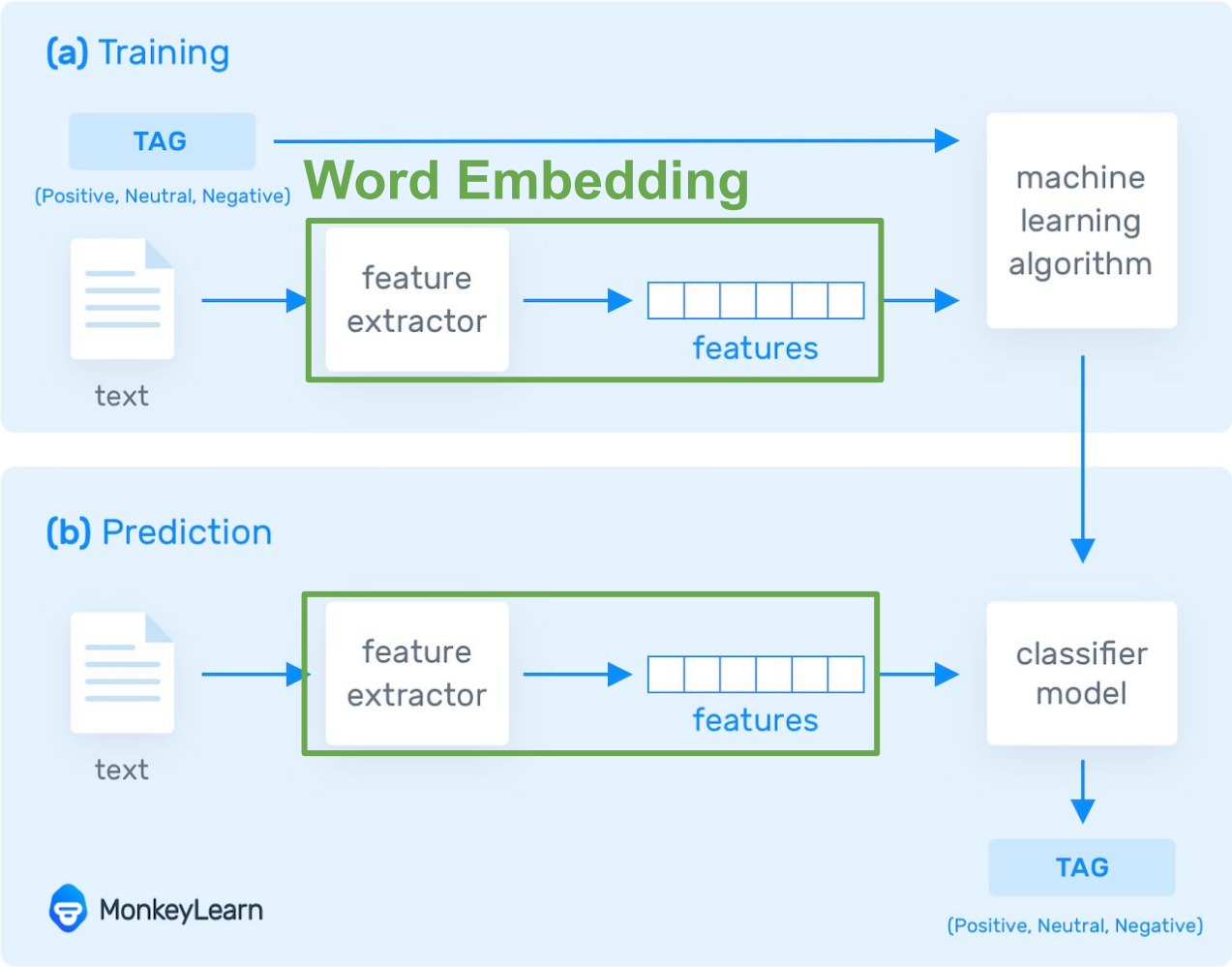
- Ambiguity
- Semantics
- Discourse-level Conversations between AI and User
- AI Anthropomorphism
- Language-specific problems



https://medium.com/@jrodthoughts/ambiguity-in-natural-language-processing-15f3e4706a9a

Natural Language Understanding

task: classifying



General Framework

- identify the task
- scope
- design the solution
- code
- test and/or analyze
- *deploy*