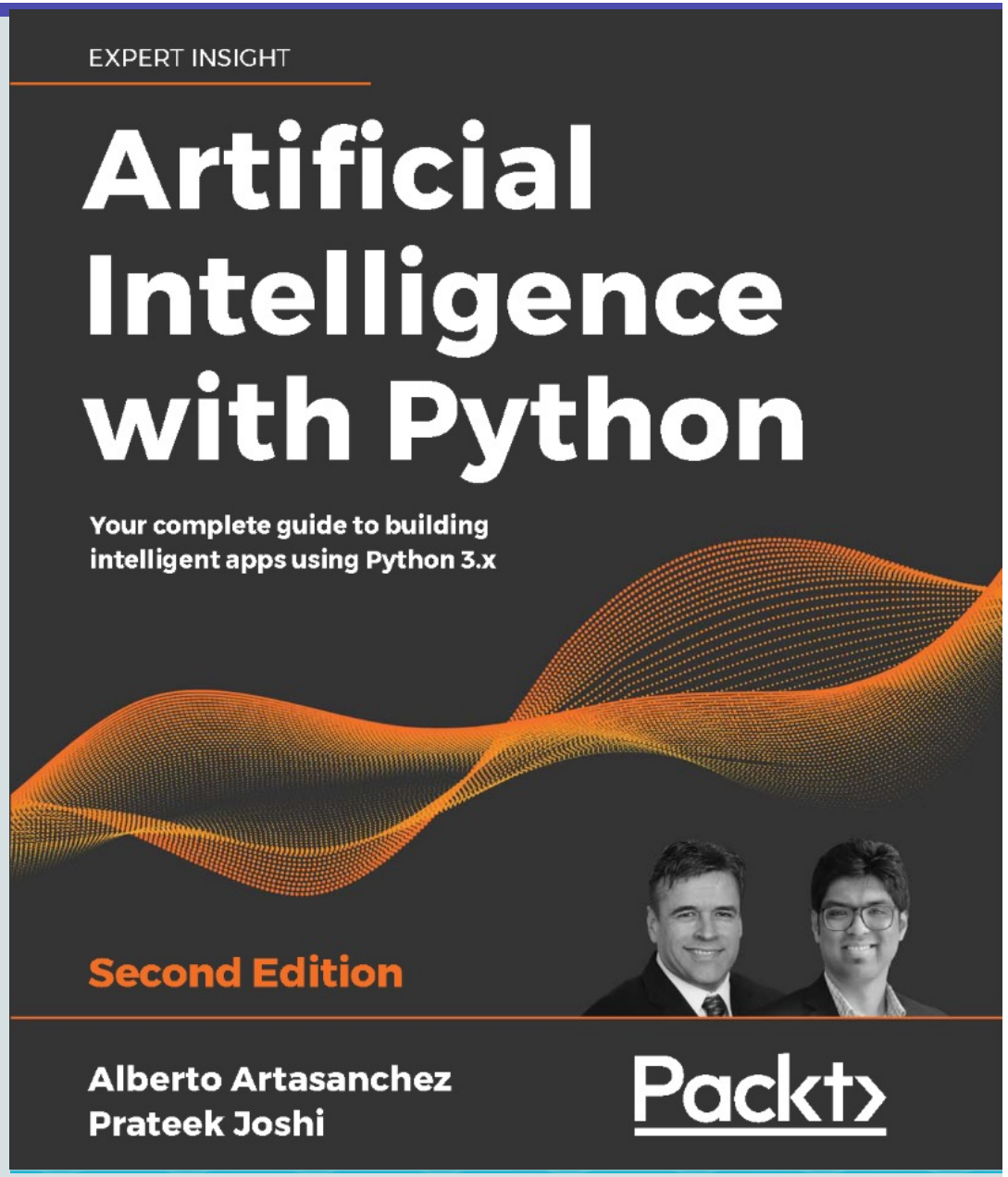
University of South-Eastern Norway
School of Business

Lecture 6: Natural Language Processing

Sinuo Wu

Course: AI for Business Applications (AI3000)



Quiz

- Enter room number or Scan the QR code

Today

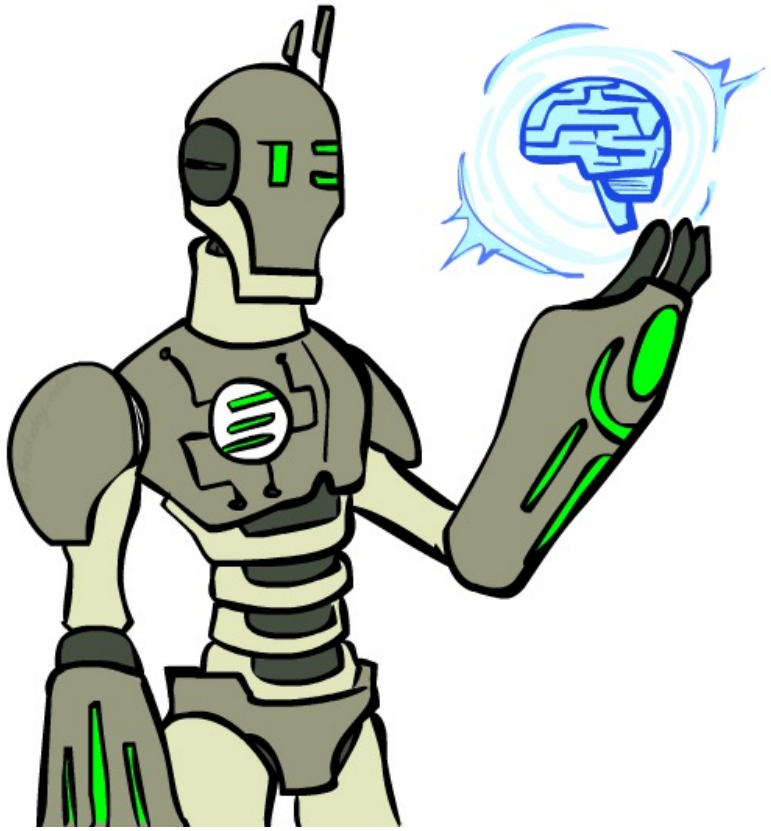
- **Natural language processing (NLP)**
- **Text Preprocess**

- Tokenization

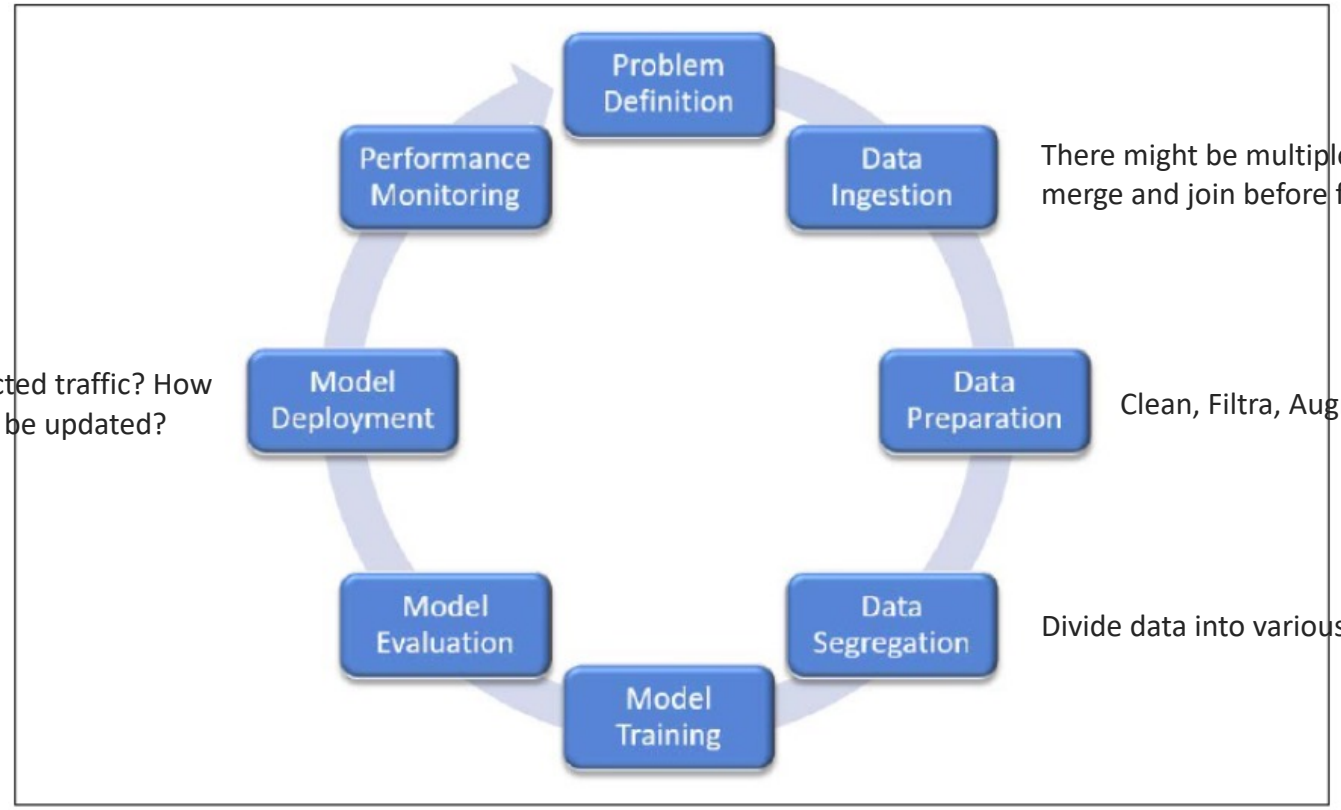
- Stemming

- Lemmatization

- **Sentiment Analysis**
- **Text Transformation**
- **Applications of NLP in business**



Steps in a ML pipeline



There might be multiple sources that we need to merge and join before feeding them to the model.

Clean, Filtra, Augment, etc.

Divide data into various categories for difference purpose

Real-time? Expected traffic? How often it needs to be updated?

Figure 1: The machine learning pipeline

Natural Language Processing

Natural Language Processing (NLP)

- NLU(Understanding) + NLG (Generation)
- **Goal:** To have computers able to understand human language
- A speaker (or writer) has the **goal** of communicating some knowledge, then plans some language that represents the **knowledge**, and **acts** to achieve the goal. The listener (or reader) **perceives** the language and **infers** the intended meaning.

NLP

Structured data well

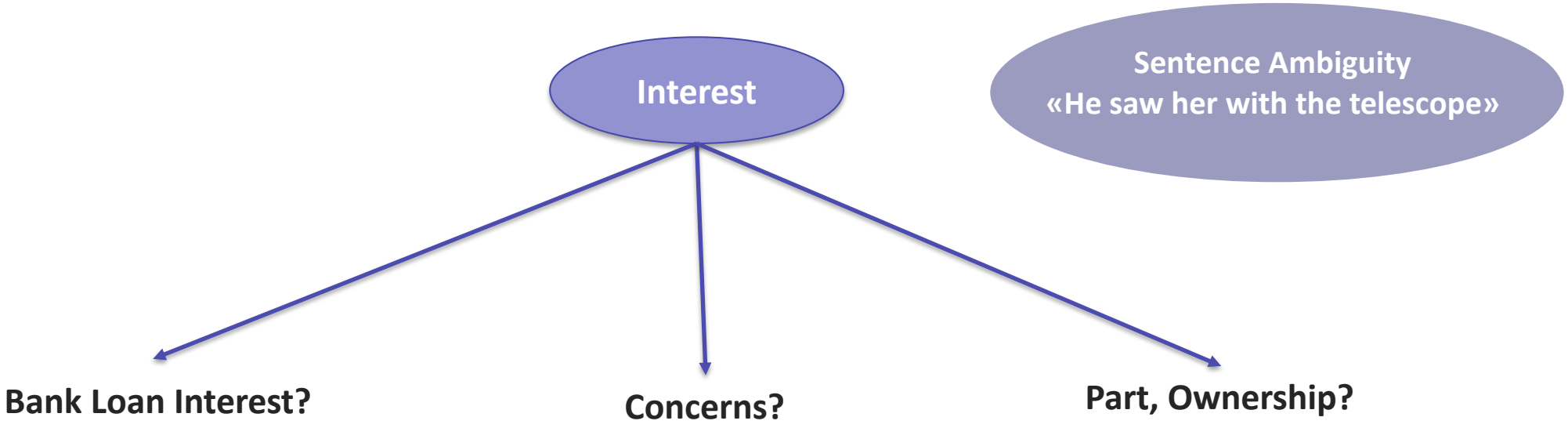
staffNo	fName	lName	salary
SL21	John	White	30000
SG37	Ann	Beech	12000
SG14	David	Ford	18000
SA9	Mary	Howe	9000
SG5	Susan	Brand	24000
SL41	Julie	Lee	9000

Free-form text

It is used extensively in search engines, conversational interfaces, document processors, and so on. Machines can handle structured data well, but when it comes to working with free-form text, they have a hard time. The goal of NLP is to **develop algorithms that enable computers to understand free - form text** and help them understand language.

NLP

- Ambiguity



NLP

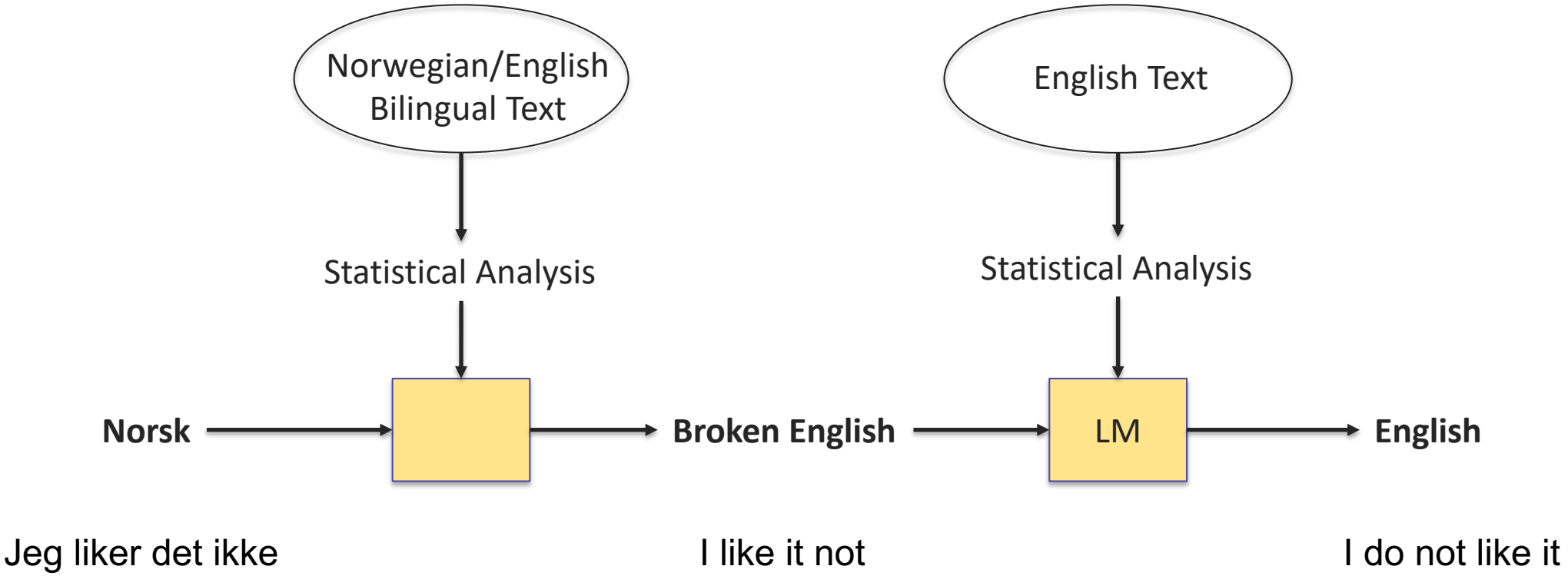
- Multiple ways to express

«Not to be invited is sad» P(Not to be invited is sad)

«To be not invited is sad» P(To be not invited is sad)

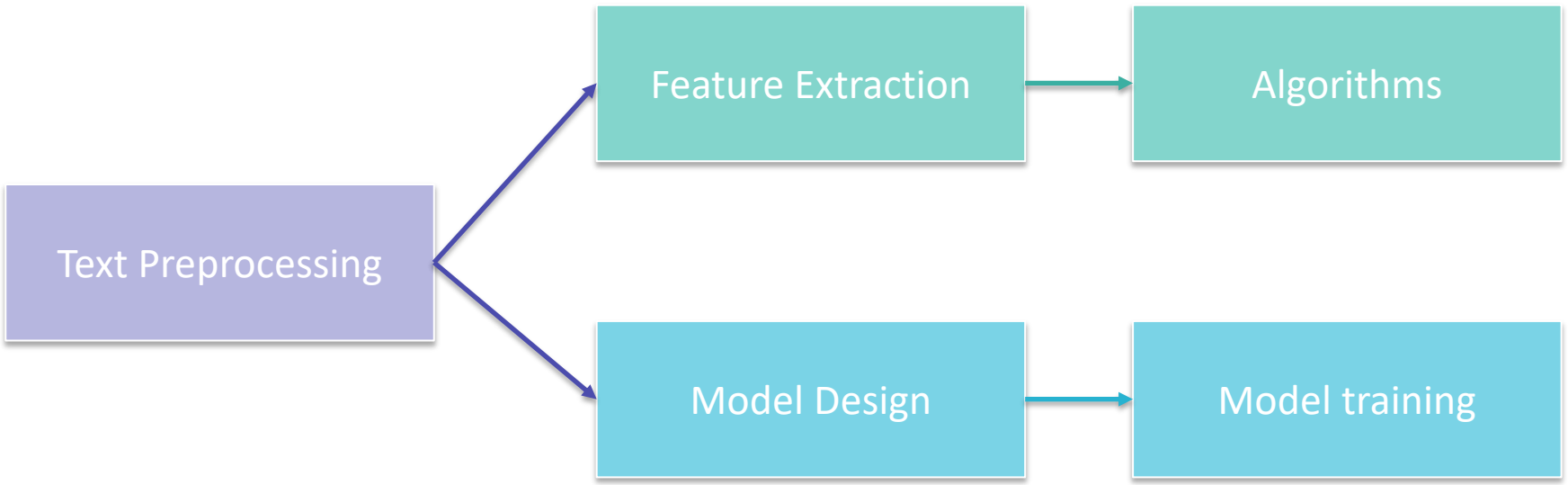
What is the probability that a random sentence would be applied?
P(S=words)

NLP


Norwegian/English Bilingual Text
English Text
Statistical Analysis
Statistical Analysis
Norsk
Broken English
LM
English
Jeg liker det ikke
I like it not
I do not like it

NLP

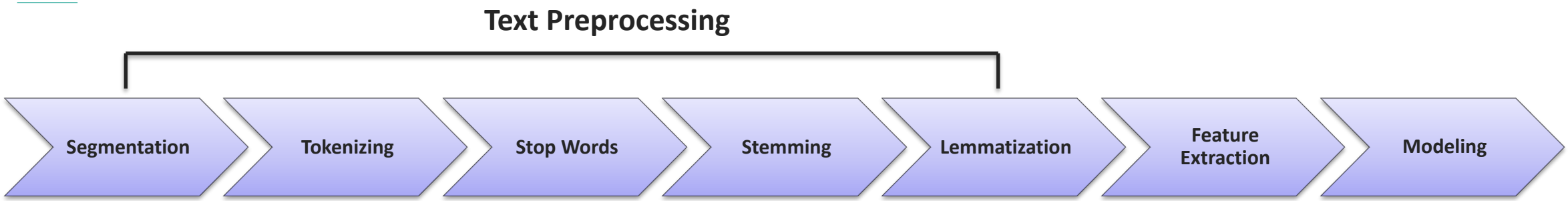
How it works in ML?



Traditional
Machine Learning

Deep Learning

NLP Pipeline



1. Segmentation <<< break data into sentences.
2. Tokenizing <<< break sentence into words (Some people also use Tokenizing to represent the first two steps)
3. Stop Words <<< mark down "the," "and" "is," "in," "it," and "an.", ...etc...'
4. Stemming <<< reduce the word to its base form.
5. Lemmatization <<< find the base or dictionary form of words.
6. Feature Extraction <<< Transform text into numerical data.
7. Modeling <<< Use algorithms to build the model (ex. naive bayes classification).

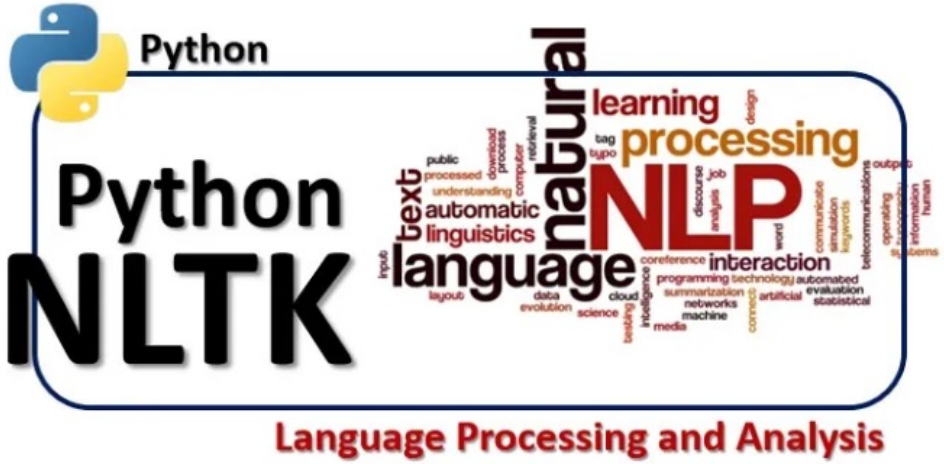
NLTK

- NLTK (Toolkit): A popular open-source library for natural language processing (NLP) in Python

Text preprocessing: NLTK can be used to preprocess text data before it is fed into ML models. This can include tasks like tokenization, stemming, and stop words removal.

Sentiment analysis: NLTK can be used to perform sentiment analysis on text data, allowing you to determine whether a piece of text express a positive, negative, or neutral emotion.

Text classification: NLTK can be used to classify text data into categories, for example, spam or non-spam emails, positive or negative reviews.



Tokenization

- Segmentation & Tokenization

```
Sentence tokenizer:
['Do you know how tokenization works?', "It's actually quite interesting!", "Let's analyze a couple of se
ntences and figure it out."]

Word tokenizer:
['Do', 'you', 'know', 'how', 'tokenization', 'works', '?', 'It', "'s", 'actually', 'quite', 'interesting'
, '!', 'Let', "'s", 'analyze', 'a', 'couple', 'of', 'sentences', 'and', 'figure', 'it', 'out', '.']

Word punct tokenizer:
['Do', 'you', 'know', 'how', 'tokenization', 'works', '?', 'It', "'", 's', 'actually', 'quite', 'interest
ing', '!', 'Let', "'", 's', 'analyze', 'a', 'couple', 'of', 'sentences', 'and', 'figure', 'it', 'out', '.
']
```

Do not mix: *While tokenization is well known for its use in cybersecurity and in the creation of NFTs, tokenization is also an important part of the NLP process. Tokenization is used in natural language processing to split paragraphs and sentences into smaller units that can be more easily assigned meaning.*

Tokenization

- Try it!

The NLTK library provides various tokenization options, and "Punkt" is one of the tokenizers available within NLTK.

```
import nltk
nltk.download('punkt')  # Download the necessary data for tokenization

from nltk.tokenize import sent_tokenize

# Sample sentences for tokenization
sample_text = "Hello, Mr. Smith. How are you today? I hope you're doing well. Have a great day!"

# Tokenize the sentences
sentences = sent_tokenize(sample_text)

# Print the tokenized sentences
for sentence in sentences:
    print(sentence)
```

Stemming

- Stemming: Reduce a word to its base word

INPUT WORD	PORTER	LANCASTER	SNOWBALL
writing	write	writ	write
calves	calv	calv	calv
be	be	be	be
branded	brand	brand	brand
horse	hors	hors	hors
randomize	random	random	random
possibly	possibl	poss	possibl
provision	provis	provid	provis
hospital	hospit	hospit	hospit
kept	kept	kept	kept
scratchy	scratchi	scratchy	scratchi
code	code	cod	code

Porter, Lancaster, and Snowball are stemming algorithms in NLP.

•**Porter**: Basic and widely used, follows simple rules.

•**Lancaster**: More aggressive, shorter stems.

•**Snowball**: An improved Porter, supports multiple languages.

Stemming

- Try it!

```
import nltk
from nltk.stem import SnowballStemmer

# Initialize the Snowball stemmer for English
snowball_stemmer = SnowballStemmer("english")

# List of words to stem
words = ["running", "eating", "adjustable", "jumped", "jumps"]

# Stem each word and print the results
stemmed_words = [snowball_stemmer.stem(word) for word in words]

print("Original words:", words)
print("Stemmed words:", stemmed_words)
```

Original words:
['running', 'eating', 'adjustable', 'jumped', 'jumps’]

Stemmed words:
['run', 'eat', 'adjust', 'jump', 'jump']

Lemmatization

- Lemmatization: Find the base or dictionary form of words.

INPUT WORD	NOUN LEMMATIZER	VERB LEMMATIZER
writing	writing	write
calves	calf	calve
be	be	be
branded	branded	brand
horse	horse	horse
randomize	randomize	randomize
possibly	possibly	possibly
provision	provision	provision
hospital	hospital	hospital
kept	kept	keep
scratchy	scratchy	scratchy
code	code	code

Lemmatization plays a crucial role in NLP tasks like text classification, information retrieval, and sentiment analysis, where understanding **the base form of words** can simplify text analysis and improve accuracy.

A "Noun Lemmatizer" and a "Verb Lemmatizer" are specific types of lemmatizers that focus on lemmatizing nouns and verbs, respectively.

Lemmatization

- Try it!

```
import nltk
from nltk.stem import WordNetLemmatizer

# Initialize the WordNet lemmatizer for verb lemmatization
verb_lemmatizer = WordNetLemmatizer()

# List of words for verb lemmatization
words = ['gone', 'ate', 'liked', 'been', 'branded', 'writing']

# Download the "wordnet" resource if not already downloaded
nltk.download('wordnet')

# Lemmatize each word as verbs and print the results
lemmatized_verbs = [verb_lemmatizer.lemmatize(word, pos='v') for word in words]

print("Original words:", words)
print("Lemmatized verbs:", lemmatized_verbs)
```

Original words:
['gone', 'ate', 'liked', 'been', 'branded', 'writing']

Lemmatized verbs:
['go', 'eat', 'like', 'be', 'brand', 'write']

Stemming VS Lemmatization

Stemming:

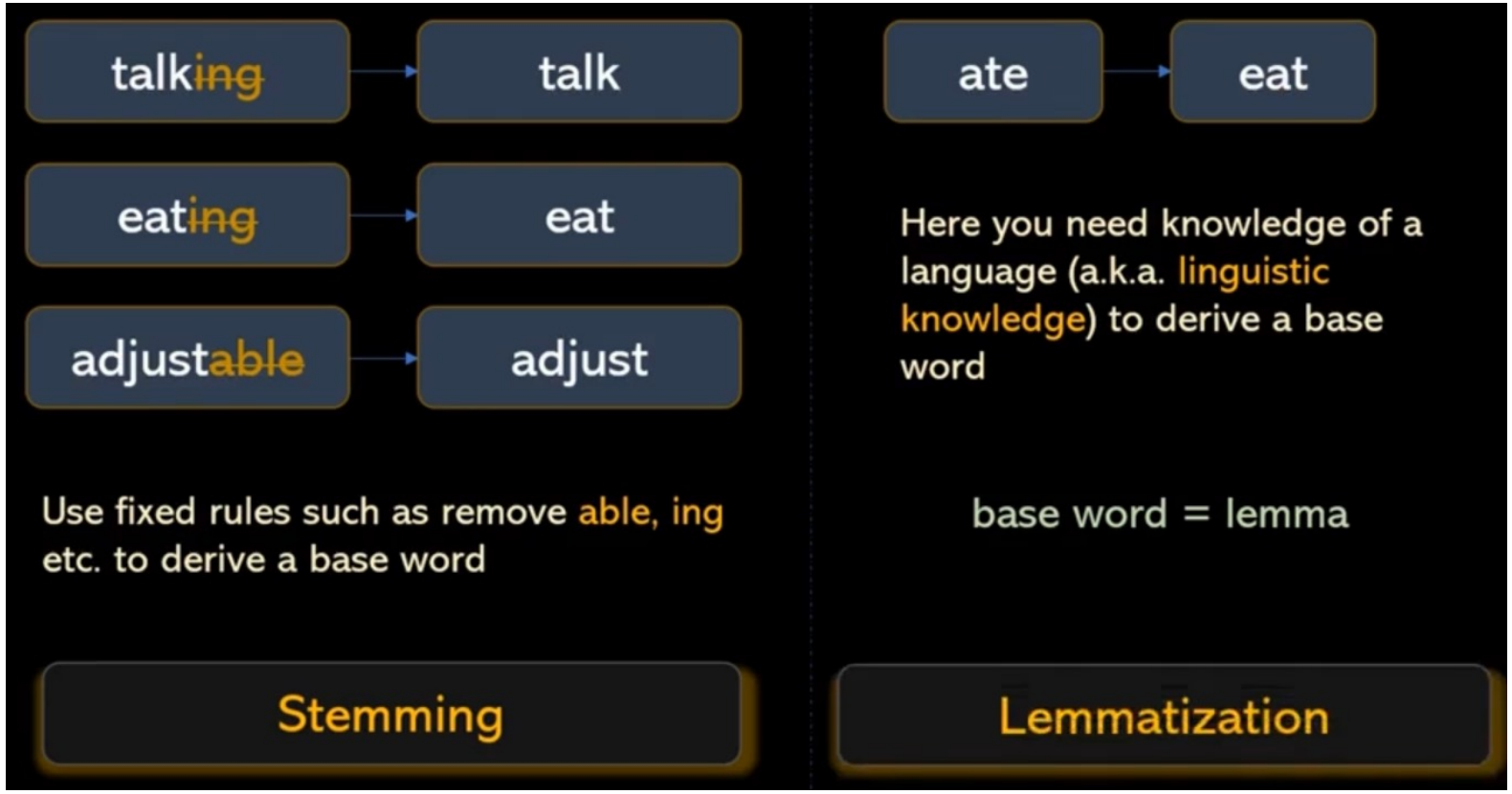
Reduce to a base word

(Simply remove)

Lemmatization:

Adjust to a base word

(Grammar!)



Have a Break!

Sentiment Analysis

Sentiment Analysis

- Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique that involves determine the sentiment or emotional tone expressed in a piece of text, such as a review, tweet, or comment.
- The primary goal of sentiment analysis is to understand and classify the sentiment as positive, negative, or neutral.

Sentiment Analysis

1. **Text Input**: Sentiment analysis takes text as input, which can be a single sentence, a paragraph, or an entire document.
2. **Text Processing**: The text is preprocessed to remove noise and irrelevant information, including punctuation, stop words, and special characters. It involves tokenization, stemming, or lemmatization to deal with the text.
3. **Sentiment Classification**: Sentiment analysis algorithms then analyze the processed text to determine the underlying sentiment. The most common sentiment classes are:

   Positive: Indicates a positive or favorable sentiment, such as happiness, satisfaction, or approval.

   Negative: Indicates a negative or unfavorable sentiment, such as anger, disappointment, or criticism.

   Neutral: Represents a lack of strong sentiment, typically indicating a neutral or objective statement.
4. **Algorithm Selection**: Various techniques can be used for sentiment analysis, ranging from rule-based methods to machine learning and deep learning approaches.
5. **Output**: The sentiment analysis model provides an output that categorizes the sentiment expressed in the text, such as "positive," "negative," or "neutral." Some models may also provide the sentiment score associated with the classification.

Applications

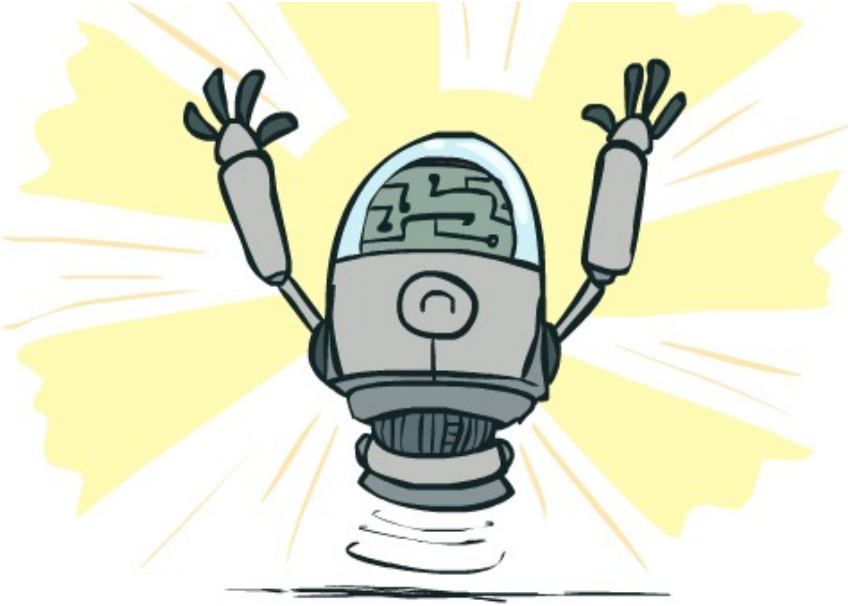
1. **Product Reviews:** Understanding customer sentiment from online reviews.
2. **Social Media Monitoring:** Analyzing public opinion on platforms.
3. **Customer Feedback:** Measuring satisfaction and finding improvements.
4. **Brand Tracking:** Monitoring a brand's reputation.
5. **Market Research:** Gathering insights on trends.
6. **Financial Analysis:** Making investment decisions based on sentiment in news and social media.

Example

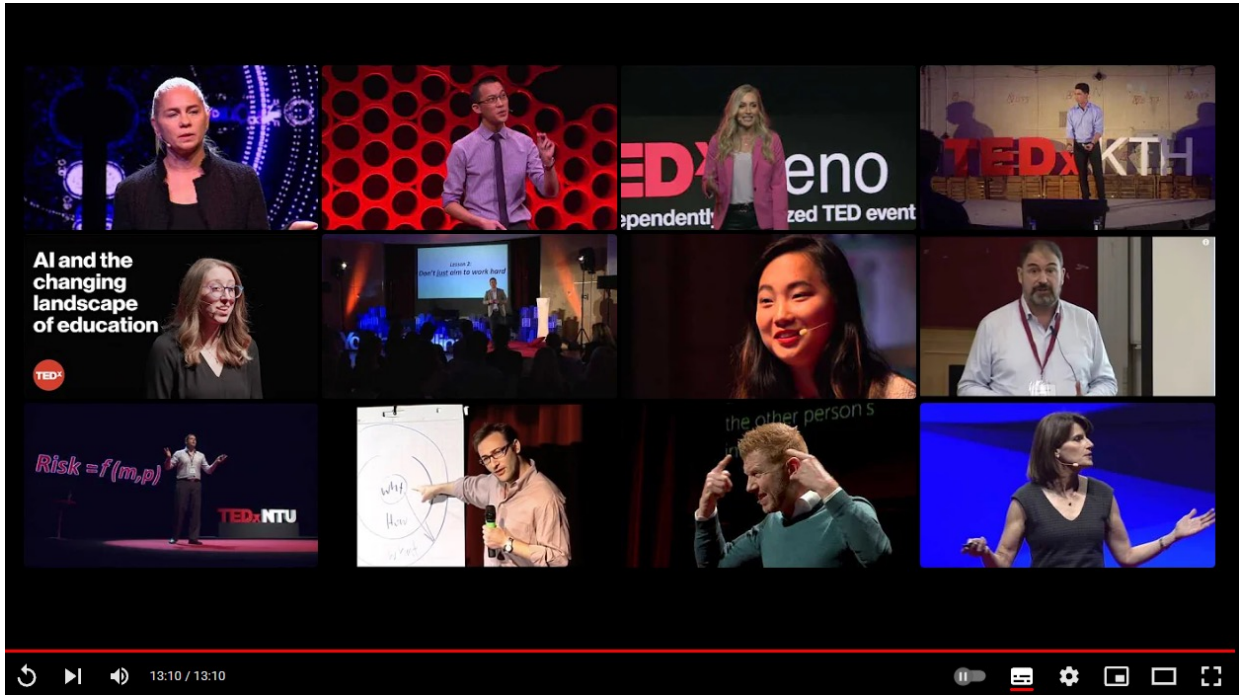
Source from Analytics Vidhya

Practice

Develop a sentiment analyzer



Inspiration



https://www.youtube.com/watch?v=n4L5hHFcGVk

Data

	A	B
1	Review	Rating
2	nice hotel expensive parking got good deal stay hotel anniversary, arrived late evening took advice previous reviews did valet parking, check quick easy, little disappointed non-existent view room room clean nice size, bed comfortable woke stiff neck high pillows, not soundproof like heard	4
3	ok nothing special charge diamond member hilton decided chain shot 20th anniversary seattle, start booked suite paid extra website description not, suite bedroom bathroom standard hotel room, took printed reservation desk showed said things like tv couch ect desk clerk told oh mixed su	2
4	nice rooms not 4* experience hotel monaco seattle good hotel n't 4* level.positives large bathroom mediterranean suite comfortable bed pillowsattentive housekeeping staffnegatives ac unit malfunctioned stay desk disorganized, missed 3 separate wakeup calls, concierge busy hard touch	3
5	unique, great stay, wonderful time hotel monaco, location excellent short stroll main downtown shopping area, pet friendly room showed no signs animal hair smells, monaco suite sleeping area big striped curtains pulled closed nice touch felt cosy, goldfish named brandi enjoyed, did n't pa	5
6	great stay great stay, went seahawk game awesome, downfall view building did n't complain, room huge staff helpful, booked hotels website seahawk package, no charge parking got voucher taxi, problem taxi driver did n't want accept voucher barely spoke english, funny thing speak arabic	5
7	love monaco staff husband stayed hotel crazy weekend attending memorial service best friend husband celebrating 12th wedding anniversary, talk mixed emotions, booked suite hotel monte carlos, loaned beautiful fan-tanned goldfish named joliet weekend visited dogs worked desk hum	5
8	cozy stay rainy city, husband spent 7 nights monaco early january 2008. business trip chance come ride.we booked monte carlo suite proved comfortable longish stay, room 905 located street building, street noise not problem view interesting rooms building look dank alley midsection large	5
9	excellent staff, housekeeping quality hotel chocked staff make feel home, experienced exceptional service desk staff concierge door men maid service needs work, maid failed tuck sheets foot bed instance soiled sheets used, staff quickley resolved soiled sheets issue, guess relates emplo	4
10	hotel stayed hotel monaco cruise, rooms generous decorated uniquely, hotel remodeled pacific bell building charm sturdiness, everytime walked bell men felt like coming home, secure, great single travelers, location fabulous, walk things pike market space needle.little grocery/drug store	5
11	excellent stayed hotel monaco past w/e delight, reception staff friendly professional room smart comfortable bed, particularly liked reception small dog received staff guests spoke loved, mild negative distance uphill ppmarket restaurants 1st, overall great experience,	5
12	poor value stayed monaco seattle july, nice hotel priced 100- 150 night not, hotel takes beating quotient, experience simply average, nothing exceptional paying 300+ n't ca n't terribly disappointed, wife stayed nicest suites 200/night felt like overpaying, best advice shop, quality-wise league	2
13	nice value seattle stayed 4 nights late 2007. looked comparable hilton marriott westin area points/miles n't gave monaco shot, pleasantly surprised nice room service quick tasty bed especially comfortable unlike people nothing positive things say service, downside relatively small exercise	4
14	nice hotel good location hotel kimpton design whimsical vibe fun, staff young casual problem hotel busy stay friendly helpful, group reserved rooms gave connecting rooms fuss, not busy week.the rooms decent size not huge pretty average city hotel, beds nice bathrooms clean bright pocke	4
15	nice hotel not nice staff hotel lovely staff quite rude, bellhop desk clerk going way make things difficult, waited forever check heavy bags no help getting throught double doors room, worst desk clerk checking tap tap tap car waited pick 20 minutes, got home noticed 1,000 dollar hold credit	3
16	great hotel night quick business trip, loved little touches like goldfish leopard print robe, complaint wifi complimentary not internet access business center, great location library service fabulous,	4
17	horrible customer service hotel stay february 3rd 4th 2007my friend picked hotel monaco appealing website online package included champagne late checkout 3 free valet gift spa weekend, friend checked room hours earlier came later, pulled valet young man just stood, asked valet open sa	1
18	disappointed say anticipating stay hotel monaco based reviews seen tripadvisor, definitely disppointment, decor room hotel envisioned nice, housekeeping staff impressive extremely polite cheery helpful, desk bellmen standard customer service, lots little things easily overlooked broken	2
19	fantastic stay monaco seattle hotel monaco holds high standards kimpton hotel line, having stayed kimpton hotels cities easily say seattle hotel monaco best seen, service attentive prompt, based member kimpton loyalty program upgraded 10th floor room great view 4th street excellent pu	5
20	good choice hotel recommended sister, great location room nice, comfortable bed- quiet- staff helpful recommendations restaurants, pike market 4 block walk, stay,	5
21	hmmmmm say really high hopes hotel monaco chose base girlfriend shopping trip seattle, stay say given competition seattle just okay, hotel lot nice features little things detract bedding super soft luxurious comforters pillows, mattress rock hard standard hotel bed kept awake nights, furnis	3
22	service service service spent week g-friend labor day bumbershoot, gray line airporter drops corner hotel 10 person cab 28 total make sure flat rate town car 38. location central downtown street w. it__Ã‡_Â©_ quick walk points tourist business, yes space needle bit far monorail westlake cer	5
23	excellent stay, delightful surprise stay monaco, thoroughly enjoyed stay, room comfortable lovely amenities friendly staff, especially enjoyed hour indulgence, definitely come,	5
24	good value downtown hotel monaco seattle great option pricey area town, rooms w street going close 400 night, hotel monaco unique interesting cozy, bed/linens quality aveda products bathroom nice touch.you close downtown quick cab ride away nightlife restaurants loved lola, experien	4
25	hotel monaco great location service hotel monaco centrally located provides excellent service, recently stayed 4/23-5/1, originally checked slight problem informed booked queen beds expedia room king bed, fortunately fellow guest checking time swapped rooms, future reference booking	5
26	great location need internally upgrade advantage north end downtown seattle great restaurants nearby good prices, rooms need updated literally thought sleeping 1970 bed old pillows sheets, net result bad nights sleep, stay location, staff friendly,	2
27	n't mind noise place great, read reviews noise used hotel website book room, interesting site gives checkbox options 1 room farthest elevator 2 room higher floor 3 quiet room, getting quietest room possible checked hoping, got 6th floor 18-19 floors hall elevator listened trucks ambulance	3
28	loved, stayed warwick overnight getway enjoy christmas shopping, warwick exceeded expectations, staff wonderful extrememly friendly room clean service lounge wonderful, came contact hotel friendly, women bathroom lever lounge well.. think haunted totally creepy vibe lights anywh	4
29	met expectations centrally located hotel blocks water popular nightlife shopping options belltown downtown, classify property star location, paid wedding rate not sure fares value proposition, room spacious little outdated, rooms decent views lower floor drab panaroma parking lot buildin	3
30	nice hotel husband stayed warwick 4 years ago liked hotel location did not hesitate book trip.we asked received non-smoking room king bed high floor, hoping hotel nice view space needle, check-in nice friendly, keys minutes nice room small loveseat juliet balcony.love location hotel just	4
31	good hotel not large hotel newly decorated, rooms good size clean, excellent location 2 blocks centre town 4 blocks market area.hotel resturant bustling place alot diners coming street, good meal there.overall highly recommend staying,	4
32	good choice seattle stayed night business booked company, overall satisfactory arrived late evening room 17th floor clean comfortable great view, bathroom little small clean equipped, good night sleep enjoyed excellent breakfast restaurant checking, staff encountered friendly polite,	4
33	great location expensive parking warwick heart seattle easy walking distance pikes place monorail terminal downtown shops, stayed twice recently start finish tour pacific north west, staff helpful knowledgable rooms large comfy armchair large bedded room, rooms fridge coffee maker.parl	4
34	noise airconditioner-a standard, arranged stay travel agency unfortunately warwick seattle hotel dissapointment trip, 3 night stay warwick changed 3 rooms, starting minute stay hotel personnel didn`t make feel like guest like intruder, reluctant help solve complaints, hotel right downtown	1
35	good location poor cleanliness warwick hotel great location seattle, close shopping pike place seattle centre, really lets hotel cleanliness, bed linen torn curtains torn floor stains, hotel advertises having balcony reality disappointment, stayed hotel san diego did balcony table chairs, warwick	2

Process

- Load the data
- Text preprocessing (Remember the steps in ML)
- Split the data into training and testing sets
- Standardization (Transform text into numerical data)
- Train a Naive Bayes classifier
- Evaluate the model
- Save the trained model

Have a Break!

Report

One Volunteer from each group report for:

1. What is your experience
2. Show your result

Text Transformation

Vectorization

- Vectorization, in the context of data processing and machine learning, refers to the process of **converting data into numerical vectors or arrays** that can be used as input for various algorithms, including machine learning models.

- Vectorization is a fundamental concept in many fields, including natural language processing (NLP), computer vision, and data analysis. It involves representing data in a format that can be efficiently processed by computers and utilized for analytical tasks.

Vectorization

1. Julie loves John more than Linda loves John
2. Jane loves John more than Julie loves John

John	2	2
Jane	0	1
Julie	1	1
Linda	1	0
likes	0	1
loves	2	1
more	1	1
than	1	1

Two vectors

```
Item 1: [2, 0, 1, 1, 0, 2, 1, 1]
Item 2: [2, 1, 1, 0, 1, 1, 1, 1]
```

Word Vectorizer

- Methods or techniques used to convert words or text into numerical vectors.
- These numerical representations of words are used in NLP and ML tasks because most machine learning algorithms require **numerical input** rather than text data.
- Popular word vectorizers include Word2Vec, GloVe, FastText, TF-IDF (Term Frequency-Inverse Document Frequency). We just need to call them from the library.

Word Vectorizer

- Word2Vec, GloVe, FastText are not part of scikit-learn library. These models are typically implemented using external libraries and have their own dedicated Python packages for training and using word embeddings.

- TF-IDF (Term Frequency-Inverse Document Frequency) can be found from Scikit-learn for converting text documents into numerical vectors .

- Count Vectorization can be found from Scikit-learn for **converting a collection of text documents to a matrix of token counts**.

One-hot encoding example

Transfer text to statistical data

[“Life is short, I like python” ,
Life is too long, I hate python“]

Feature names: ['hate', 'is', 'life', 'like', 'long', 'python', 'short', 'too']

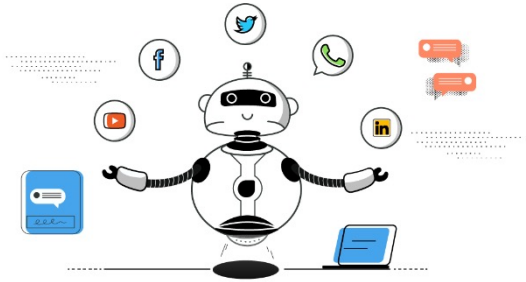
```
From sklearn.feature_extraction.text import CountVectorizer
# Input
# Create an instance
# Fit and transform the text data
# Get the feature names
# Convert the sparse matrix to an array for better visualization
# Print the feature names and the one-hot encoded representation
```

```
[[0 1 1 1 0 1 1 0]
[1 1 1 0 1 1 0 1]]
```

NLP Applications

NLP for Business

Sentiment Analysis



Chatbots

Email Filter

NLP for Business

Machine Translation

Supply Chain Support

Auto-Correct

Case Study – Email Filter

Train an email filter model with the given data

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\ws\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\ws\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
Accuracy: 0.968609865470852
              precision    recall  f1-score   support

         ham       0.97      1.00      0.98       966
        spam       1.00      0.77      0.87       149

    accuracy                           0.97      1115
   macro avg       0.98      0.88      0.92      1115
weighted avg       0.97      0.97      0.97      1115
```

Case Study – Movie Reviews

Train a movie review model with the given data

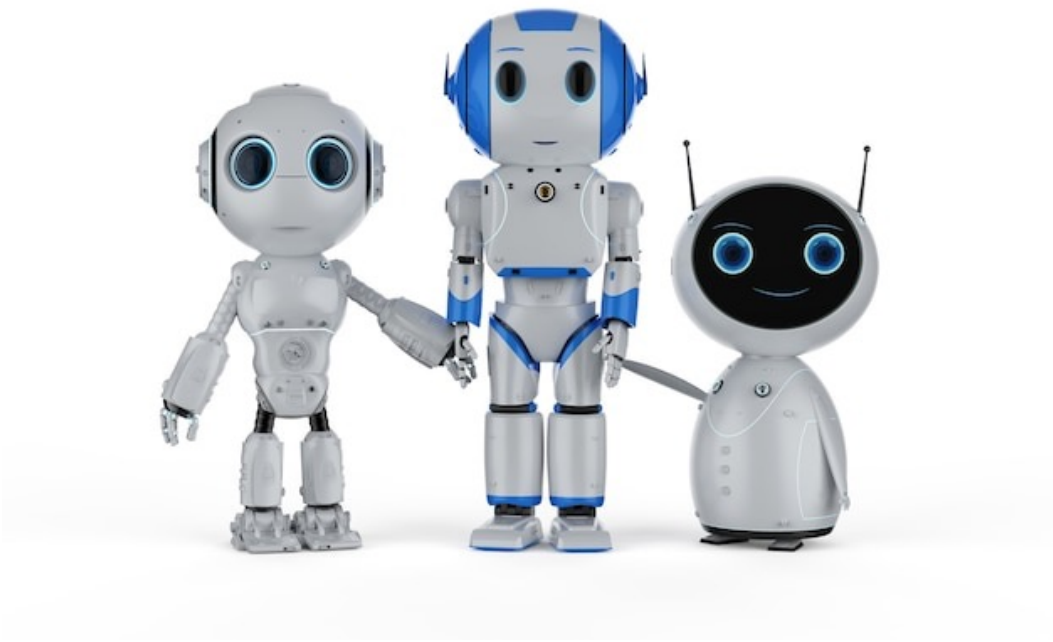
```
Accuracy: 0.5189260563380281

Classification Report:

                   precision    recall  f1-score   support

Was this review helpful?  Sign in to vote.
                      0.00      0.00      0.00        25

              1       0.50      0.01      0.02       128

             10       0.52      1.00      0.68      1178

              2       0.00      0.00      0.00        41

              3       0.00      0.00      0.00        34

              4       0.00      0.00      0.00        52

              5       0.00      0.00      0.00        61

              6       0.00      0.00      0.00        89

              7       0.00      0.00      0.00       118

              8       0.00      0.00      0.00       173

              9       0.23      0.01      0.03       373

       accuracy                           0.52      2272

      macro avg       0.11      0.09      0.07      2272

   weighted avg       0.34      0.52      0.36      2272
```

```
Top 15 most informative words:
1. Worst
2. Horrible
3. Boring
4. Rain
5. Save
6. Zero
7. indulgent
8. slapped
9. awake
10. uninteresting
11. endeavor
12. hrs
13. jumbled
14. stupidly
15. tossed

Movie review predictions:

Review: The costumes in this movie were great
Predicted sentiment: Positive
Probability: 0.92

Review: I think the story was terrible and the characters were very weak
Predicted sentiment: Positive
Probability: 0.53

Review: People say that the director of the movie is amazing
Predicted sentiment: Positive
Probability: 0.97

Review: This is such an idiotic movie. I will not recommend it to anyone.
Predicted sentiment: Positive
Probability: 0.82
```

What can you do with low accuracy?

Have a Break!

Group Up!



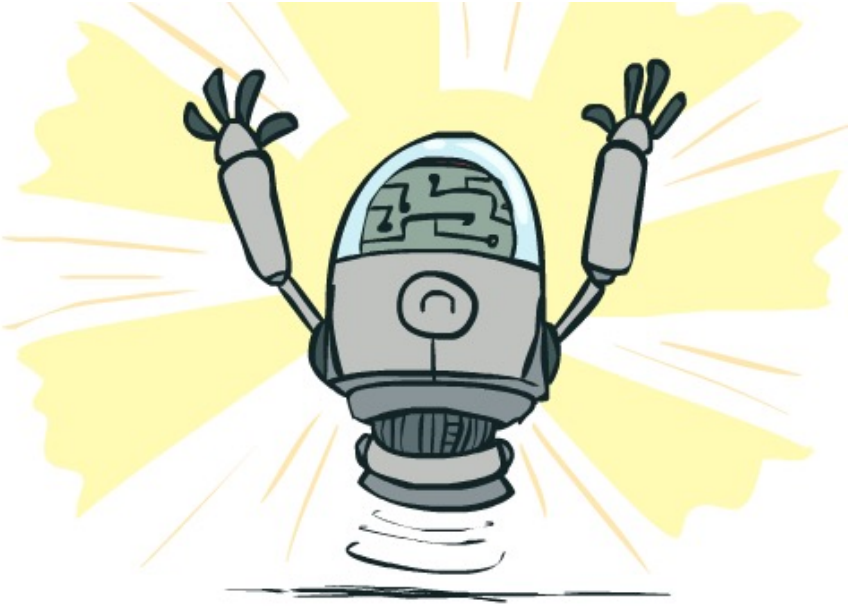
Work on your assignment!

Guidance

https://towardsdatascience.com/nlp-spam-detection-in-sms-text-data-using-deep-learning-b8632db85cc8

Reflection

ASSIGNMENT PROCESS



THANKS