NLP

NLP: Natural Language Processing

-technology used to process, analyze, and create natural language

-subfield of AI concerning interactions between computers and human language

- -Google Translate
- -chatboxes (think customer service!)
- -Amazon Echo
- -autocorrect/Grammarly

NLP: Various Languages

-NLP is not only limited to English

-some of its technology differs across languages that use different structures

-ex: character-based languages like Chinese/Japanese vs alphabet languages of English/French

-ethical concern: NLP techniques developed mainly for English

Tech Behind NLP

Naive Bayes Algorithm

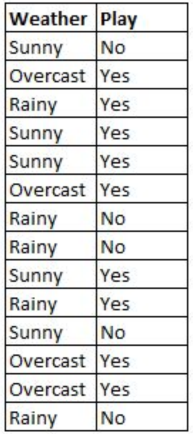
Text	Sentiment
I like apples	2
I hate apples	0
I bought apples	1

0 = negative
1 = neutral → Rated by real people
2 = positive

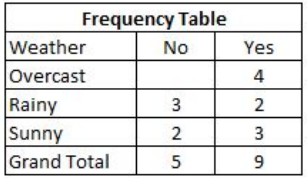
Naive Bayes Algorithm

Step 1: Convert data into frequencies

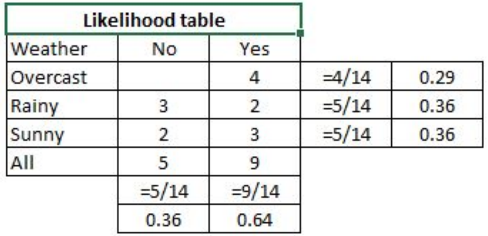
Step 2: Create likelihood table with data



Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No



Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9



Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Step 3: Use data to figure out sentiment

Naive Bayes Algorithm

Text	Sentiment
I like bananas	2
I hate cauliflowers	0
I bought some books	1

0 = negative
1 = neutral → Rated by machine
2 = positive

Naive Bayes Algorithm

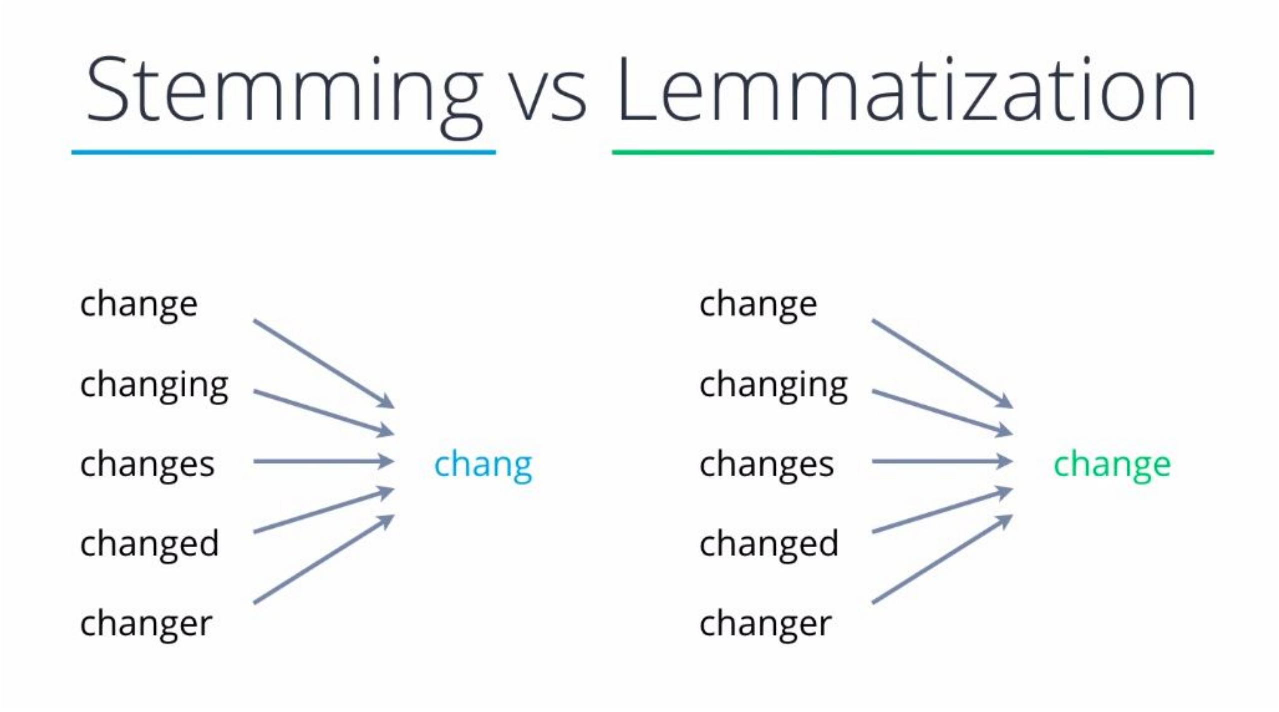
Naive Bayes is mostly used for:

1. Text classification/ Spam Filtering/ Sentiment Analysis
2. Recommendation System

However, it is also a great example of how machines process words.

Language is complex, how do we simplify it for computers?

Stemming and Lemmatization



Stemming vs Lemmatization

change
changing
changes
changed
changer
→ chang

change
changing
changes
changed
changer
→ change

Stemming and Lemmatization

Stemming is the process of reducing a word to its stem or root format.

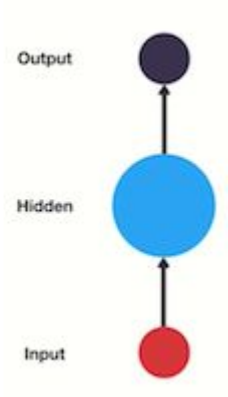
In lemmatization, the transformation uses a dictionary to map different variants of a word back to its root format.

Stemming is *easier and faster*, whereas lemmatization is more *accurate and preserves context* → hardly vs hard is a case only lemmatization can process

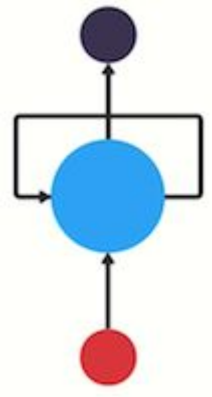
RNN

RNN, aka Recurrent Neural Networks, is a slight twist form the normal neural network.

Neural Network



RNN



RNN

Special thing about RNN is that instead of making all the words numbers like Naive Bayes, it tries to use sequential memory.

Sequential Memory:

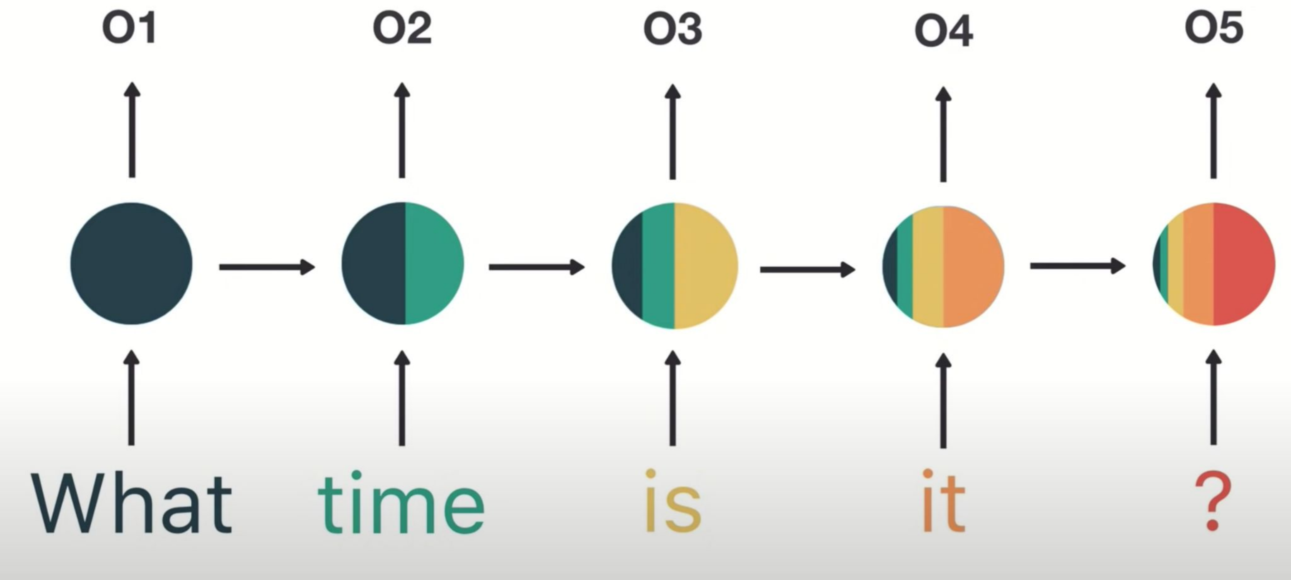
ABCDEFGHIJKLMNOPQRSTUVWXYZ

ZYXWVUTSRQPONMLKJIHGFEDCBA

NOPQRSTUVWXYZABCDEFGHIJKLM

RNN – Chatbot Example

Let’s say you ask this question, it will process each word by part like such:

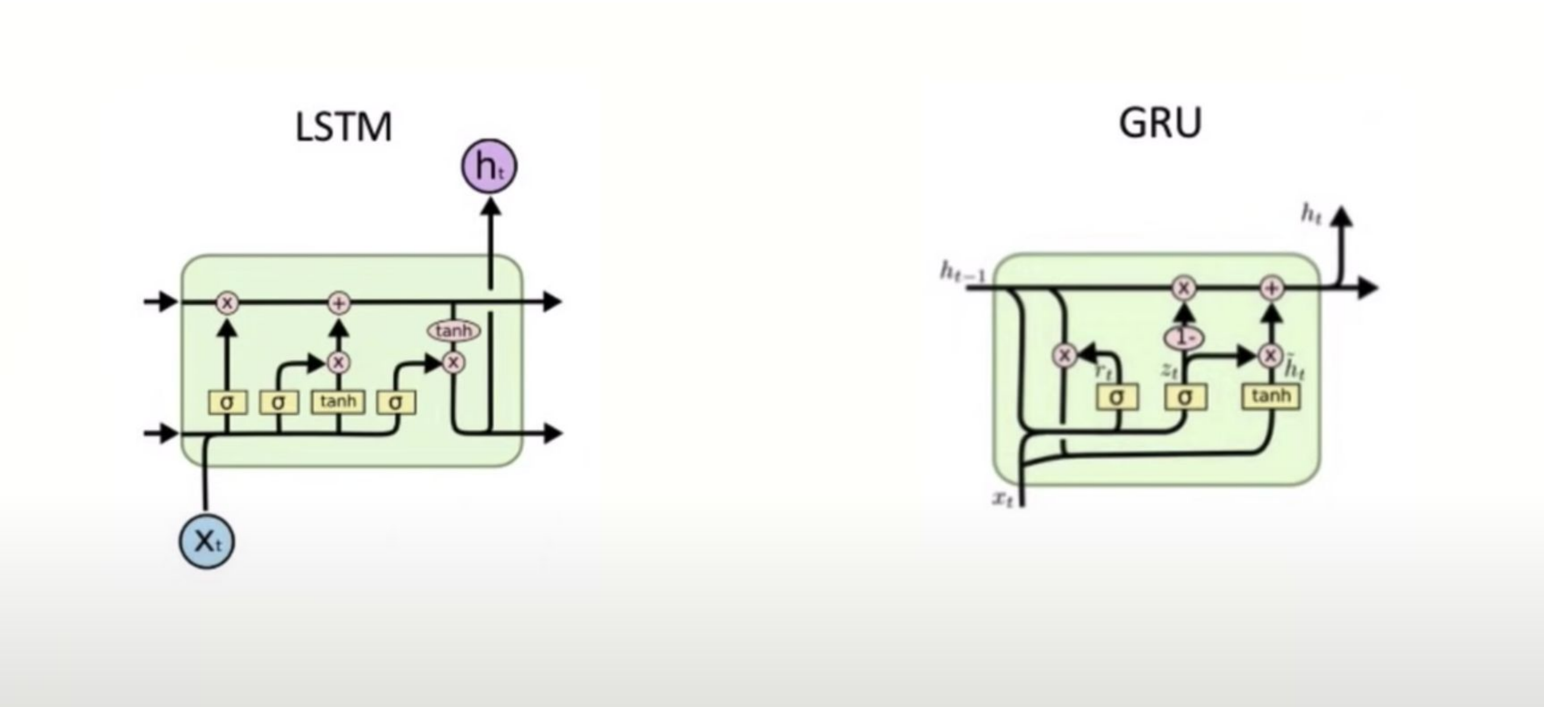


RNN

However, RNN has limited memory, therefore the gradient. As you can see, the first few words are basically out of the system now.

The lack of context is why you get funny results sometimes.

LSTM and GRU



https://youtu.be/8HyCNIVRbSU?t=180

Watch until 8:30

https://youtu.be/8HyCNIVRbSU?t=581

LSTM and GRU

The video touches on a lot of technical terms, so it is fine if you didn’t understand most of it. However, here are some key takeaways:

1. LSTM and GRU has better memory and therefore understand context a bit better
2. GRU is a simplified version of LSTM but works almost as well
3. Both LSTM and GRU are expensive and takes a long time to train

Speaking of context…

As you can tell, A.I., ML, or DL all depend on converting data into numbers, and despite the complexity of these machines, the context rarely gets carried over. The lack of understanding often leads to bias and can cause massive damage if it goes undetected.

Demonstration!

https://gpt3demo.com/apps/openai-gpt-3-playground

Discussion Questions: