CONFIDENTIAL

26 March 2019

BOT IN A DAY

REPLY
VALOREM

INTRODUCTION TO NATURAL LANGUAGE PROCESSING

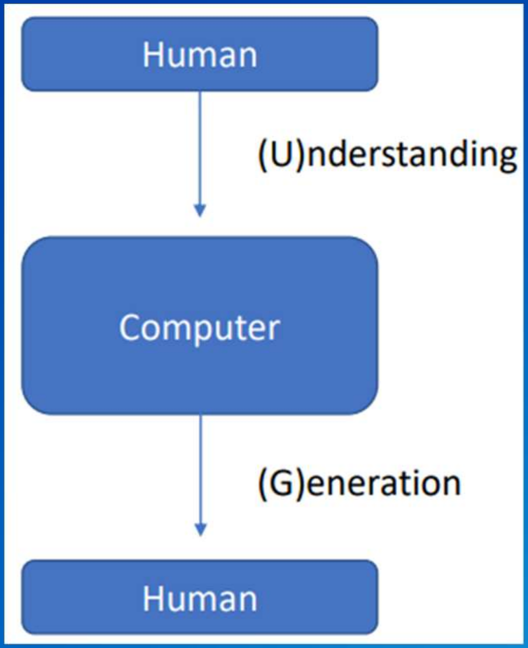
NLP Use Cases

- Understanding Intent
  - Search Engines
- Question Answering
  - Azure QnA, Bots, Watson
- Digital Assistants
  - Cortana, Siri, Alexa
- Translation Systems
  - Azure Language Translation, Google Translate
- News Digest
  - Flipboard, Facebook, Twitter
- Other uses
  - Polling, Crime mapping, Earthquake prediction

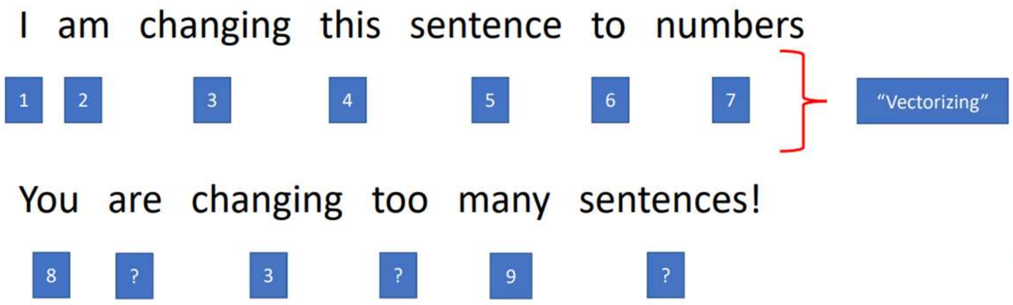
UNDERSTANDING HUMANS IS HARD

NLP requires inputs from :

- Linguistics
- Computer Science
- Mathematics
- Statistics
- Machine Learning
- Psychology
- Database Engineering



KEY:
CHANGE UNCERTAINTY TO CERTAINTY



I	am	changing	this	sentence	to	numbers
1	2	3	4	5	6	7

"Vectorizing"

You	are	changing	too	many	sentences!
8	?	3	?	9	?

Vectorizing

Mapping words to numbers to reduce ambiguity

MAPPING WORDS TO NUMBERS:
CORPUS CREATION, VECTORIZING, TFIDF

- **Corpus Creation**
  - Create a library of all words in original dataset
- **Vectorizing**
  - Changing words to numbers
  - Typically a raw count of frequency(Bag of Words)
- **TFIDF**
  - Term Frequency / Inverse Document Frequency
  - Example:
    - “This” mentioned 3 times in a given review, and the review has 27 words in it
    - Tfidf = 3 / 27 = 1/9

THE BAG OF WORDS APPROACH

- P(Positive Review | Words Contained)
- Look at the unordered words of a document to determine underlying characteristics
- Coffee reviews with the word ‘bean’ tend to be far more positive
- Common in sentiment and feature analysis

BAYES RULE

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

Example from Charles Dickens:

- P(“Darnay looked at Dr. Manette”)
- Use maximum likelihood estimates for the n-gram probabilities
  - Unigram: P(w) = c(w)/V
  - Bigram: P(w1 | w2) = c(w1,w2)/c(w2)
- Values
  - P(“Darnay”) = 533 / 598633 = .00089
  - P(“looked”|”Darnay”) = 3 / 676 = .0044
  - P(“at|looked”) = 77 / 312 = .247
  - P(“Dr. Manette” | “at”) = 2 / 4512 = .000443
- Bigram probability
  - P(“Darnay looked at Dr. Manette”) = 4.28 * e^-10
  - P(“at Dr. Manette Darnay looked”) = 0

CHALLENGES IN NLP

SYNTAX VS. SEMANTICS

Syntax

- Lamb a Mary had little
- Colorless orange liquid

Structure

- Grammatically ok but makes no sense
- Grammatically ok but makes no sense, a liquid cannot be both colorless and orange

Compiles but Meaningless

```
if 2==2:
  print(“Hello World”)
```

Semantics

- Merry hat hey lid tell lam
- I no like!

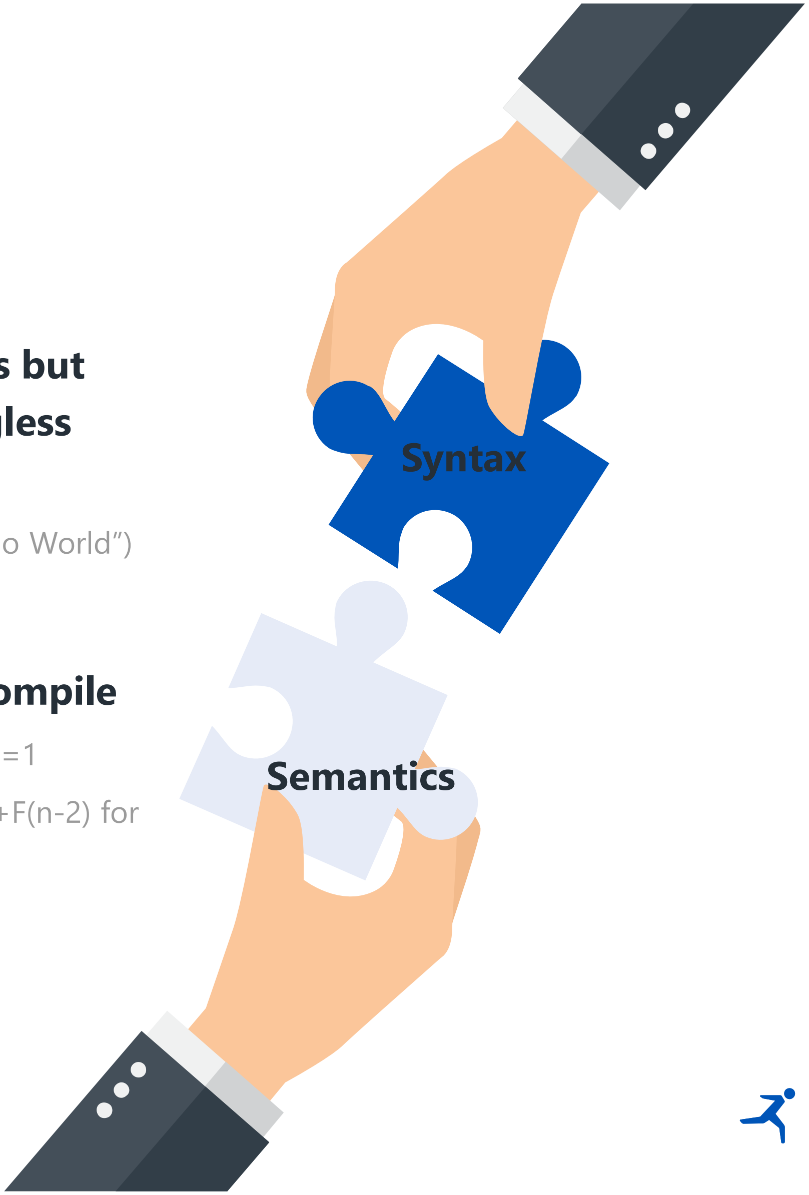
Meaning

- Has meaning but uses the wrong syntax for vocabulary
- Childlike syntax but clear semantics

Won’t Compile

F(0)=0, F(1) =1

F(n)=F(n-1)+F(n-2) for n>1



CHALLENGES IN NLP: AMBIGUITY I

Prepositional Phrase Attachment

- You ate spaghetti with meatballs / pleasure / a fork / Jillian
- Incorrectly attaching positional phrases is a large source of error in current parsing systems.

Metonymy

- Sydney is essential to this class
- Figure of speech replacing a thing or concept with the name of something closely associated

Ellipsis and Parallelism

- I gave the Steven a shovel and Joseph a ruler
- Ellipsis: omitting clauses when context is clear
- Parallelism: compounding words that have equivalent meanings

Phonetic

- My toes are getting number

CHALLENGES IN NLP: AMBIGUITY II

Referential

- Sharon complimented Lisl. She had been kind all day.

Subjectivity

- Karen believes that the Economy will stay strong

Reflexive

- Brandon brought himself an apple

Syntactic

- Call Wayne a Dentist

OTHER CHALLENGES IN NLP

Parsing N-grams

- United States of America
- Hot dog

Typos

- John Hopkins vs Johns Hopkins

Non-standard language

- (208)929-6136 vs 208-929-6136
- Cause = because

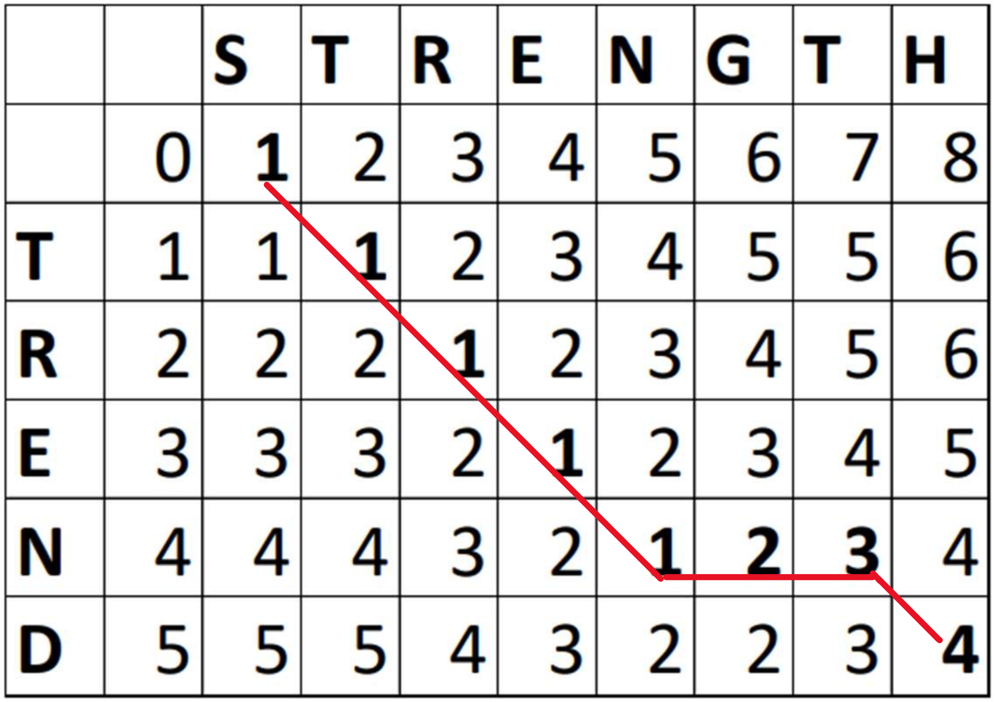
SARCASM

- Human's are *so* clear with language

HOW DO WE SPELLCHECK?

EDIT DISTANCE

- Can reference box above, left, or diagonal up-left
- If letter matches, +0
- If letter doesn't match, +1
- Score is the box at the bottom-right

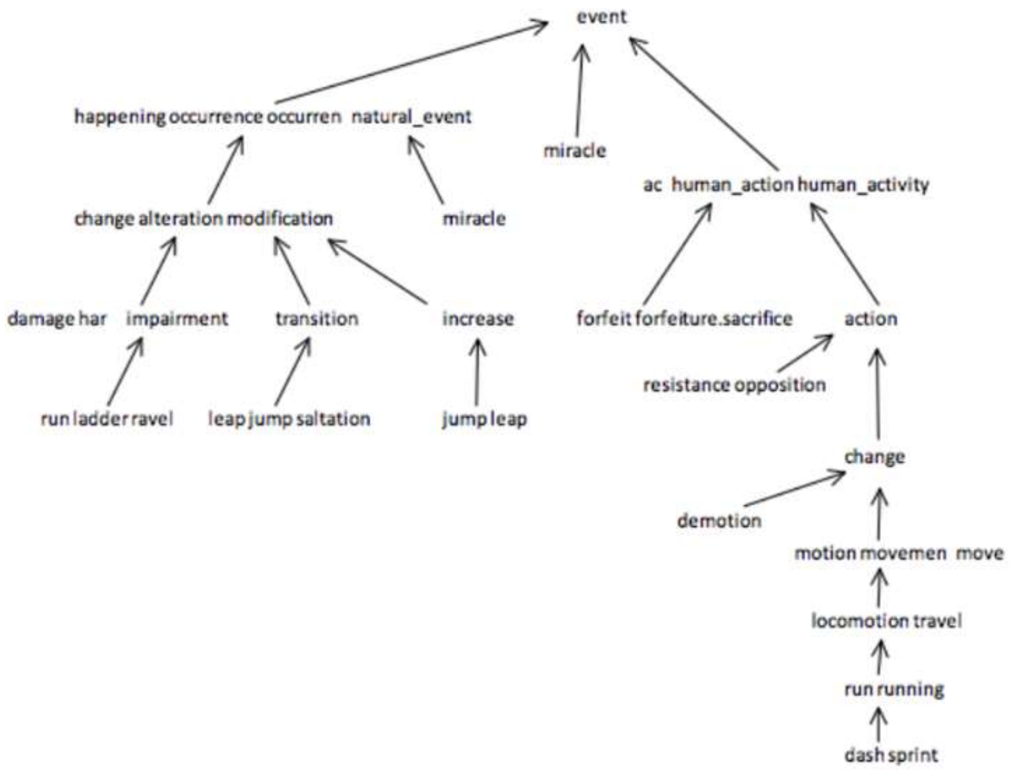


		S	T	R	E	N	G	T	H
	0	**1**	2	3	4	5	6	7	8
T	1	1	**1**	2	3	4	5	5	6
R	2	2	2	**1**	2	3	4	5	6
E	3	3	3	2	**1**	2	3	4	5
N	4	4	4	3	2	**1**	**2**	**3**	4
D	5	5	5	4	3	2	2	3	**4**

HOW DO WE DETERMINE SEMANTIC RELATIONSHIPS?

ASSOCIATIONS THAT EXIST BETWEEN THE MEANINGS OF WORDS

- Use a tree structure(Wordnet) to model relationships between words
- Measures how words are related to each other.
- Birdcage will be more like Dog Kennel than it will be to Bird
- Many different systems to draw out semantic relationships, but ‘Wordnet’ is one of the most commonly used
- Similarity metric:
  Sim(V,W) = - ln(pathlength(V,W))
  Sim(Run, Miracle) would be = -ln(7)

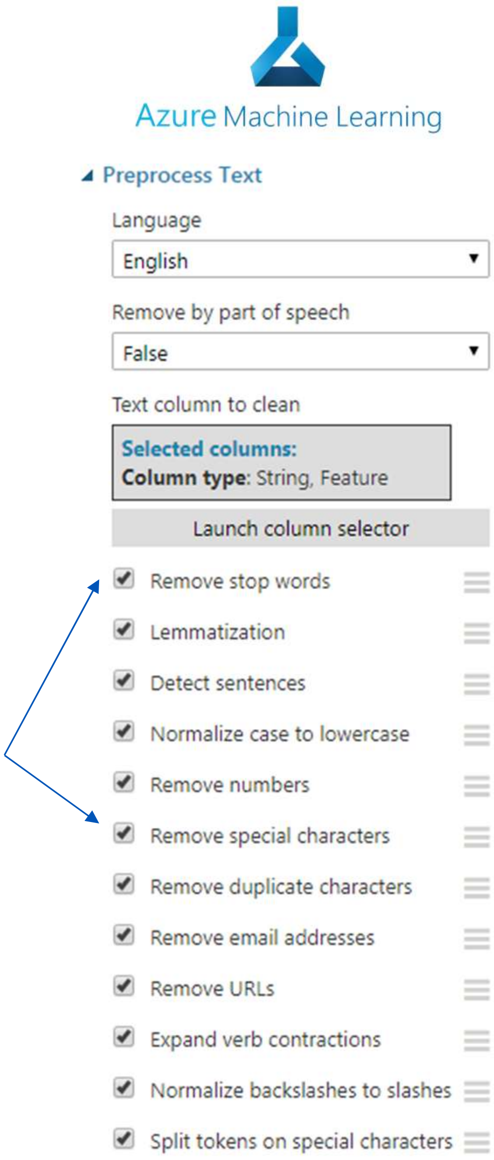


PREPROCESSING:
REMOVING STOPWORDS AND PUNCTUATION

Advantage of removing them?

- "And", "If", "But", ".", "," - Will almost ALWAYS be your most significant words
- Therefore they tell you nothing about what's going on in the text you're processing

Note: if you are focussing on Natural Language Generation you should NOT remove these



Azure Machine Learning

Preprocess Text

Language: English

Remove by part of speech: False

Text column to clean

Selected columns:
Column type: String, Feature

☑ Remove stop words
☑ Lemmatization
☑ Detect sentences
☑ Normalize case to lowercase
☑ Remove numbers
☑ Remove special characters
☑ Remove duplicate characters
☑ Remove email addresses
☑ Remove URLs
☑ Expand verb contractions
☑ Normalize backslashes to slashes
☑ Split tokens on special characters

PREPROCESSING: MEASURING AND STEMMING

Measure

- A '**measure**' of a word is an indication of how many syllables are in it.
- Consonants = 'C', Vowels = 'V'
- Every sequence of 'VC' is counted as +1
- Intellectual = (VC)C(VC)C(VC)CV(VC) = 4

Stemming: Porter's Algorithm

- Strip a word down to its barest form
- Ex: 'Alleviation' – $\underbrace{\text{'ation' + 'ate'}}_{\text{Transformation rule}}$ = 'Alleviate'
- The stem isn't always a word
- argue, argued, argues, arguing, and argus -> argu

Stemming: Sample Rules

- If measure >0:
  - Lies -> li
    - Abilities = Abiliti
  - Ational -> ate
    - National = National
    - Recreational = recreate
  - Sses -> ss
    - Sunglasses = sunglass
  - Biliti -> ble
    - Abiliti = able

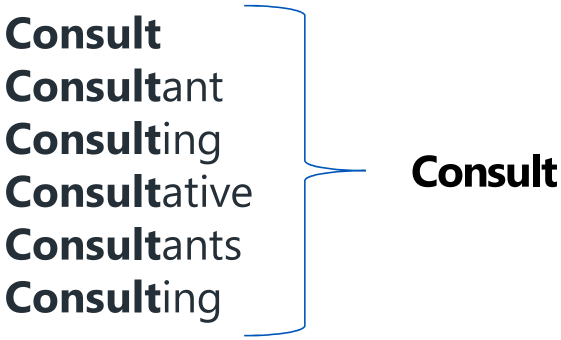
STEMMING: EXAMPLES

Computational

- Computational – ‘ational’ + ‘ate’ = Computate
- Computate – ‘ate’ = **Comput**

Computer

- Computer – ‘er’ = Comput



Consult

Consultant

Consulting

Consultative

Consultants

Consulting

Consult

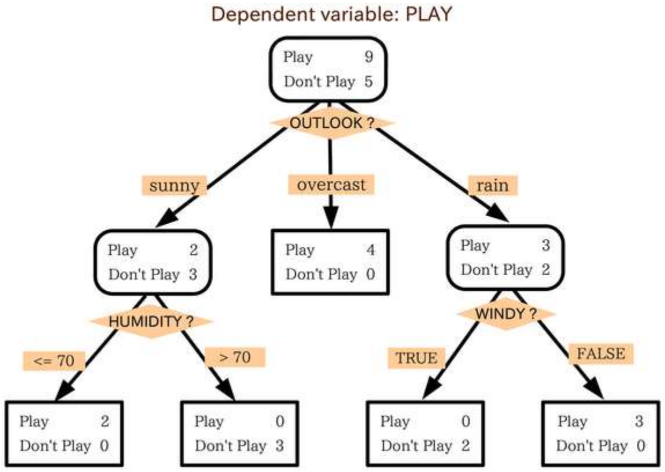
SENTENCE BOUNDARY RECOGNITION

Problem: terms like Dr., A.M., U.S.A.

Use a decision tree to estimate the boundary

Features:

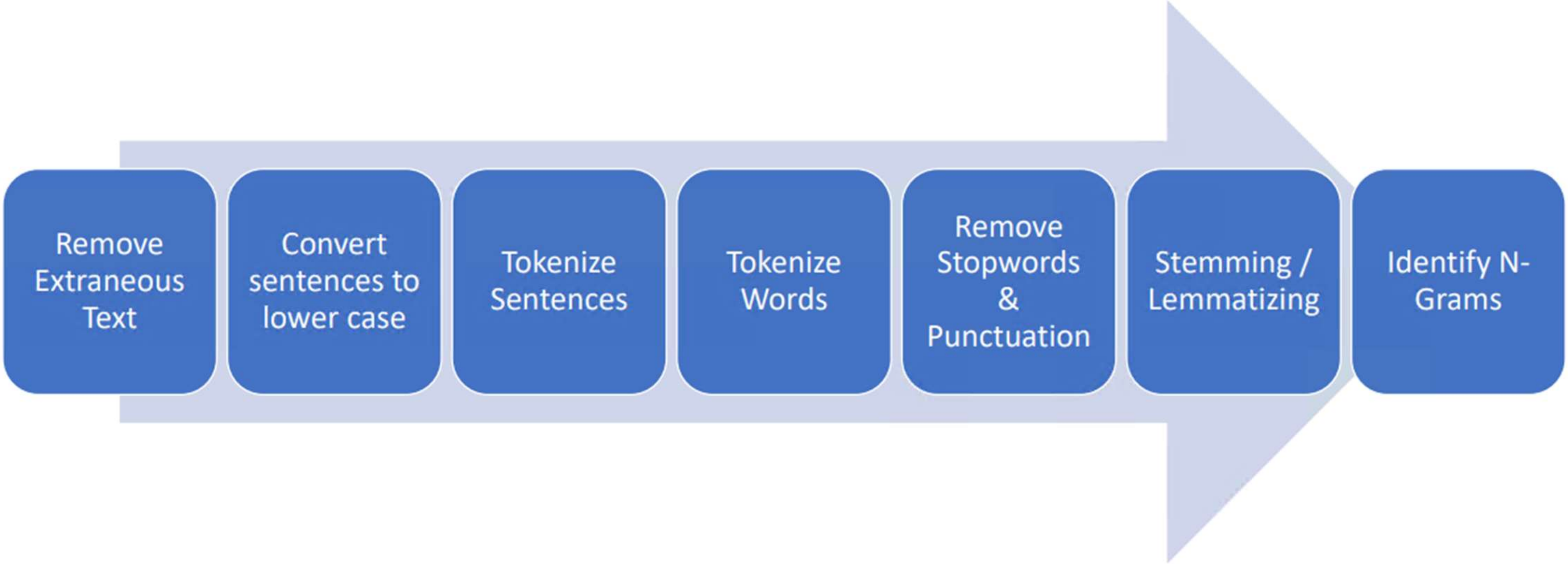
- Punctuation
- Formatting
- Fonts
- Spaces
- Capitalization
- Known Abbreviations



N-GRAM MODELING

- N-grams are words that have a distinct meaning when combined with other words
- Excellent way to highlight the importance of context when performing NLP
- Examples:
  - Unigram: Apple
  - Bigram: Hot Dog
  - Trigram: George Bush Sr.
- I'll meet you in Times _____

PRE-PROCESSING CHECKLIST


Remove Extraneous Text
Convert sentences to lower case
Tokenize Sentences
Tokenize Words
Remove Stopwords & Punctuation
Stemming / Lemmatizing
Identify N-Grams