AI Bootcamp

NLP & LLMs

Intro to NLP

Dr Usman Zia

Asst Professor

SINES, NUST

usman.zia@sines.nust.edu.pk

My Profile

NUST Graduate

MS in Computer Engg and PhD in Deep Learning

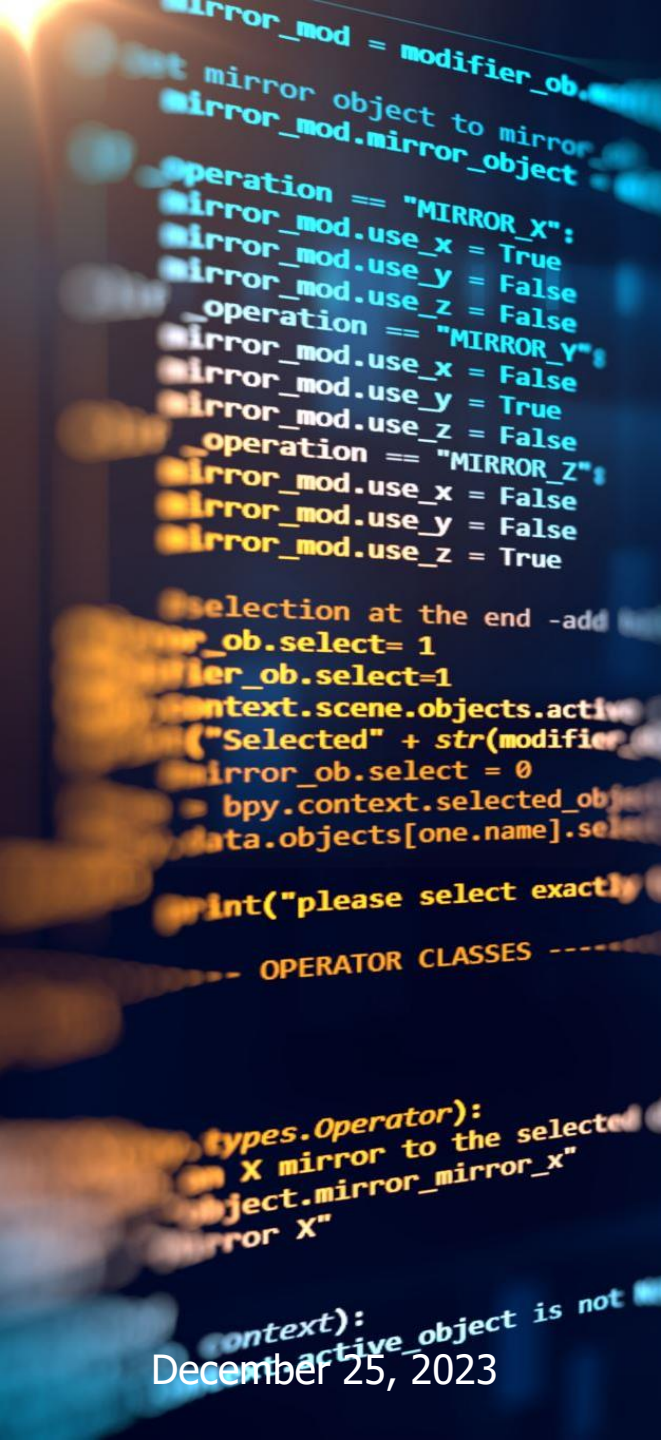
Software Developer/ Solution architect for past 18 years

Development of enterprise applications

Asst Professor (Adjunct) at SINES, NUST

Head of Jazz 5G Innovation Lab at SINES, NUST

Specialization in Language Models and Deep Learning techniques



Goals of this Field

- Computers would be a lot more useful if they could handle our email, do our library research, talk to us …
- But they are fazed by natural human language.
- How can we tell computers about language? (Or help them learn it as kids do?)

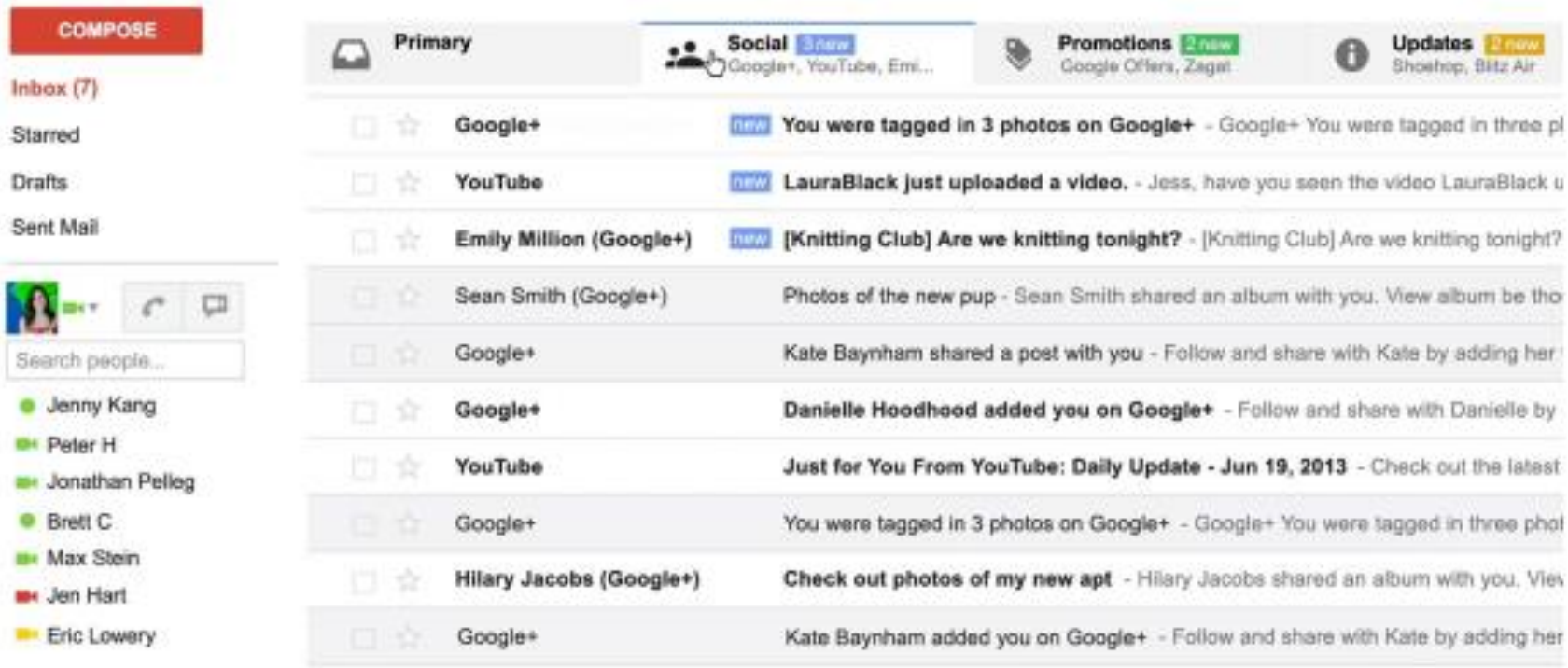
What is natural language processing?

- An experimental computer science research area that includes problems and solutions pertaining to the understanding of human language

Natural Language Processing (NLP) is the study of the computational treatment of natural (human) language.

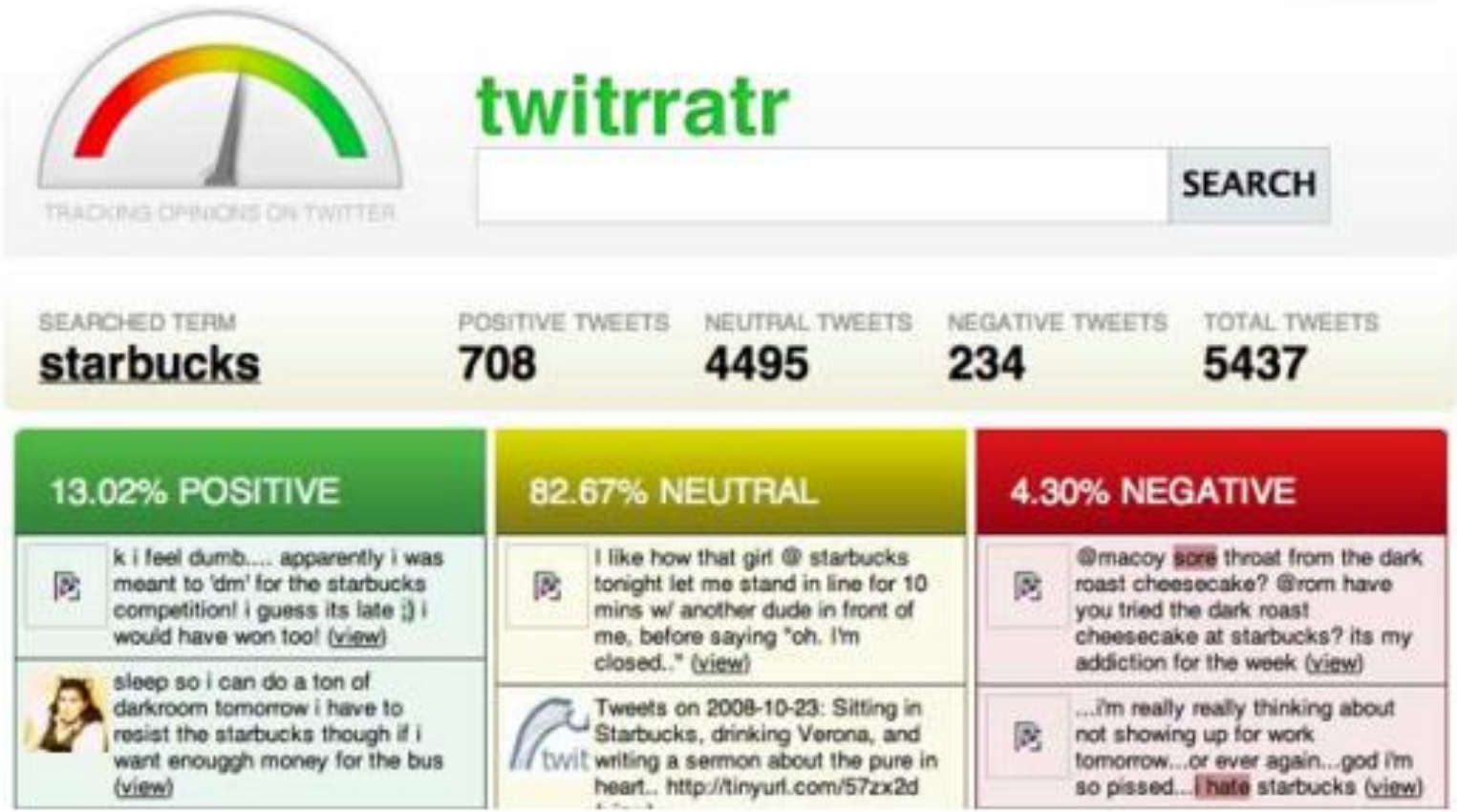
In other words, teaching computers how to understand (and generate) human language.

Text Classification



- spam / not spam
- priority level
- category (primary / social / promotions / updates)

Sentiment Analysis



Machine Translation

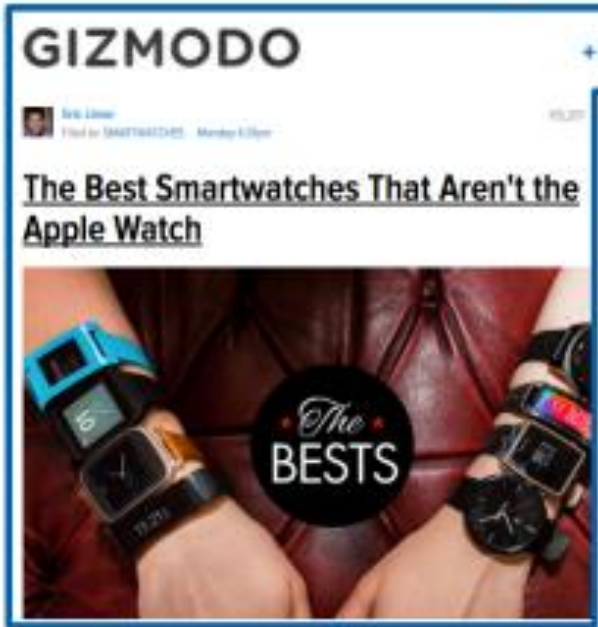
Neue Umfrage: Kaufen Sie eine Apple Watch?

New Poll: Will you buy an Apple Watch?

Question Answering

Watson had access to 200 million pages of structured and unstructured content consuming 4 TB of disk storage including the full text of the 2011 edition of Wikipedia, but was not connected to the Internet.

Summarization



The Apple Watch has drawbacks. There are other smartwatches that offer more capabilities.

Dialog Systems

user: Schedule a meeting with Matt and David on Thursday.
computer: Thursday won't work for David. How about Friday?
user: I'd prefer Monday then, but Friday would be ok if necessary.

Part-of-Speech Tagging

determiner	verb (past)	prep.	proper noun	proper noun	poss.	adj.	noun
Some	questioned	if	Tim	Cook	’s	first	product
modal	verb	det.	adjective	noun	prep.	proper noun	punc.
would	be	a	breakaway	hit	for	Apple	.

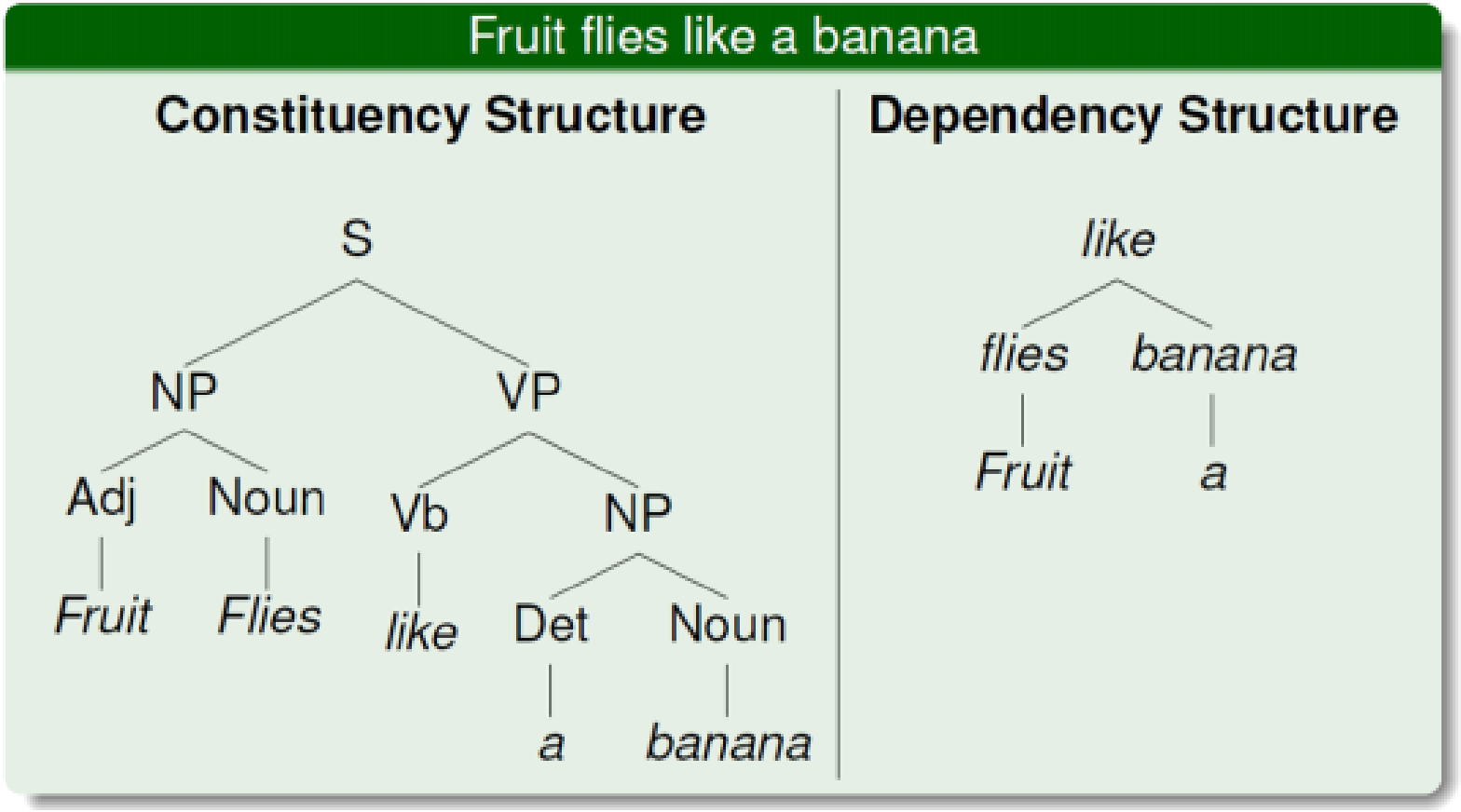
Named Entity Recognition

Some questioned if Tim Cook’s first product would be a breakaway hit for Apple.

PERSON

ORGANIZATION

Syntactic Parsing



Reading Comprehension

Once there was a boy named Fritz who loved to draw. He drew everything. In the morning, he drew a picture of his cereal with milk. His papa said, "Don't draw your cereal. Eat it!"

After school, Fritz drew a picture of his bicycle. His uncle said, "Don't draw your bicycle. Ride it!"

...

What did Fritz draw first?

A) the toothpaste

B) his mama

C) cereal and milk

D) his bicycle

Text Generation

...evelop, build and purchase the best equipment known to mankind. Our military is depleted, and we have to do it. We cannot let that happen. We're not going to happen. And yet that's a very tough night for a little while and then you have never did a deal. He did one deal. A house. And if he wins totally in favor of Common Core. He's very, very low energy. So low energy. So low energy. So low energy person, very, very low energy that everybody was apologizing to me. They saw that I want is common sense, above your safety, and above all else. I refuse to be political media – I love that sign. Look at what's going to run. He's just having fun. Just a good time. His brand – " Like I care about my brand. At this point, my brand – " Like I care about my brand. They're not so stupidly and foolishly gave them.

INSTA-TRUMP

Conspicuous by their absence…

- speech recognition (see TTIC 31110)
- information retrieval and web search
- knowledge representation
- recommender systems

Modern applications

- Search engines
- Natural language assistants
- …..
- …..
- …..
- Generative AI

Natural language Processing

- Automating the analysis, generation, and acquisition of human ("natural") language
  - **Analysis** (or "understanding" or "processing" ...)
  - **Generation**
  - **Acquisition**

Component of NLP

- **Natural Language Understanding**
  - Taking some spoken/typed sentence and working out what it means
- **Natural Language 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)

Why NLP is Hard?

- Ambiguity and variability of linguistic expression:
  - variability: many forms can mean the same thing
  - ambiguity: one form can mean many things

    I saw the boy on the beach with my binoculars

- There are many different kinds of ambiguity

  I reached the bank after crossing the _____.
  river? road?

- Each NLP task has to address a distinct set of kinds

Why NLP is Hard?

Christopher Robin is alive and well. **He** is the same person that you read about in the book, **Winnie the Pooh.** As **a boy**, **Chris** lived in a pretty home called **Cotchfield Farm**. When **Chris** was three years old, **his father** wrote a poem about **him**. The poem was printed in a magazine for others to read. **Mr. Robin** then wrote a book

- Who wrote **Winnie the Pooh?**
- Where did **Chris** live?

Why NLP is Hard?

Ambiguity

credit: A. Zwicky

Why NLP is Hard?

- Word sense / meaning ambiguity

Why NLP is Hard?

Ambiguity

San Jose cops kill man with knife

Text Paper Translate Listen Close

San Jose cops kill man with knife

Ex-college football player, 23, shot 9 times allegedly charged police at fiancee's home

By Hamed Aleaziz and Vivian Ho

A man fatally shot by San Jose police officers while allegedly charging at them with a knife was a 23-year-old former football player at De Anza College in Cupertino who was distraught and depressed, his family said Thursday.

Police officials said two officers opened fire Wednesday afternoon on Phillip Watkins outside his fiancee's home because they feared for their lives. The officers had been drawn to the home, officials said, by a 911 call reporting an armed home invasion that, it turned out, had been made by Watkins himself.

But the mother of Watkins' fiancee, who also lives in the home on the 1300 block of Sherman Street, said she witnessed the shooting and described it as excessive. Faye Buchanan said the confrontation happened shortly after she called a suicide intervention hotline in hopes of getting Watkins medical help.

Watkins' 911 call came in at 5:01 p.m., said Sgt. Heather Randol, a San Jose police spokeswoman. "The caller stated there was a male breaking into his home armed with a knife," Randol said. "The caller also stated he was locked in an upstairs bedroom with his children and requested help from police."

She said Watkins was on the sidewalk in front of the home when two officers got there. He was holding a knife with a 4-inch blade and ran toward the officers in a threatening manner, Randol said.

"Both officers ordered the suspect to stop and drop the knife," Randol said. "The suspect continued to charge the officers with the knife in his hand. Both officers, fearing for their safety and defense of their life, fired at the suspect."

On the police radio, one officer said, "We have a male with a knife. He's walking toward us."

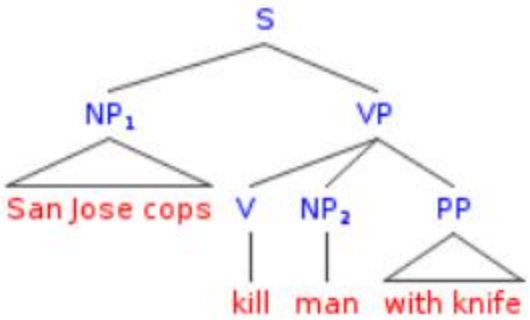
"Shots fired! Shots fired!" an officer said moments later.

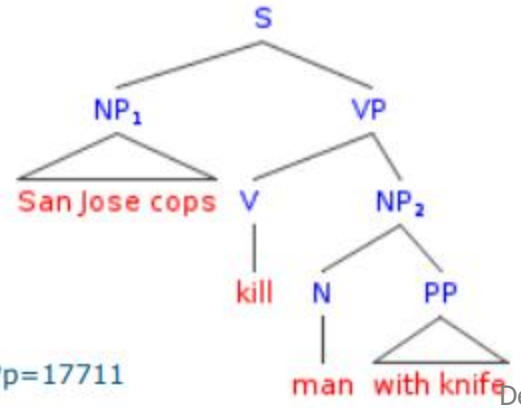
A short time later, an officer reported, "Male is down. Knife's still in hand."

Buchanan said she had been prompted to call the

Shoot continues on D8

Back Continue





Credit: Mark Liberman, http://languagelog.ldc.upenn.edu/nll/?p=17711

Why NLP is Hard?

Ambiguity

- Ambiguous headlines:
  - Include your children when baking cookies
  - Local High School Dropouts Cut in Half
  - Hospitals are Sued by 7 Foot Doctors
  - Iraqi Head Seeks Arms

  - Safety Experts Say School Bus Passengers Should Be Belted
  - Teacher Strikes Idle Kids

Why NLP is Hard?

Language is dynamic

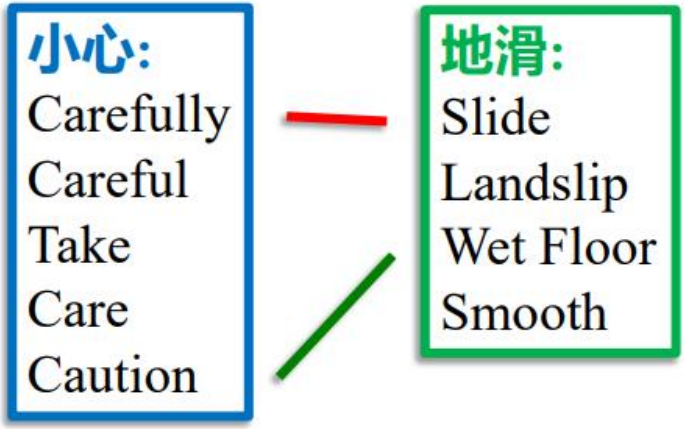
LOL	Laugh out loud
G2G	Got to go
BFN	Bye for now
B4N	Bye for now
Idk	I don't know
FWIW	For what it's worth

Why NLP is Hard?

Language is Compositional

Why NLP is Hard?

Language is Compositional



Why NLP is Hard?

Scale

- Examples:
  - Bible (King James version): ~700K
  - Penn Tree bank ~1M from Wall street journal
  - Newswire collection: 500M+
  - Wikipedia: 2.9 billion word (English)
  - Web: several billions of words

STEPS of NLP

- Morphological and Lexical Analysis
- Syntactic Analysis
- Semantic Analysis
- Discourse Integration
- Pragmatic Analysis

STEPS of NLP

- Morphology: What is a word?
- **奧林匹克運動會**（希臘語：**Ολυμπιακοί Αγώνες**，簡稱**奧運會**或**奧運**）是國際奧林匹克委員會主辦的包含多種體育運動項目的國際性運動會，每四年舉行一次。
- كبيوتها = “to her houses”

- Lexicography: What does each word mean?
  - He plays bass guitar.
  - That bass was delicious!

- Syntax: How do the words relate to each other?
  - The dog bit the man. ≠ The man bit the dog.
  - But in Russian: человек собаку съел = человек съел собаку

STEPS of NLP

- Semantics: How can we infer meaning from sentences?
  - I saw the man on the hill with the telescope.
  - The ipod is so small! ☺
  - The monitor is so small! ☹
- Discourse: How about across many sentences?
  - President Bush met with President-Elect Obama today at the White House. He welcomed him, and showed him around.
  - Who is "he"? Who is "him"? How would a computer figure that out?

AI Bootcamp

NLP & LLMs

Text Pre-Processing

usman.zia@sines.nust.edu.pk linkedin.com/in/usmanxia

Words

- *they lay back on the San Francisco grass and looked at the stars and their*

- ***Token**: an instance of that type in running text.*
- *How many?*
  - *15 tokens (or 14)*

How many words?

N = *number of tokens*

V = *vocabulary*

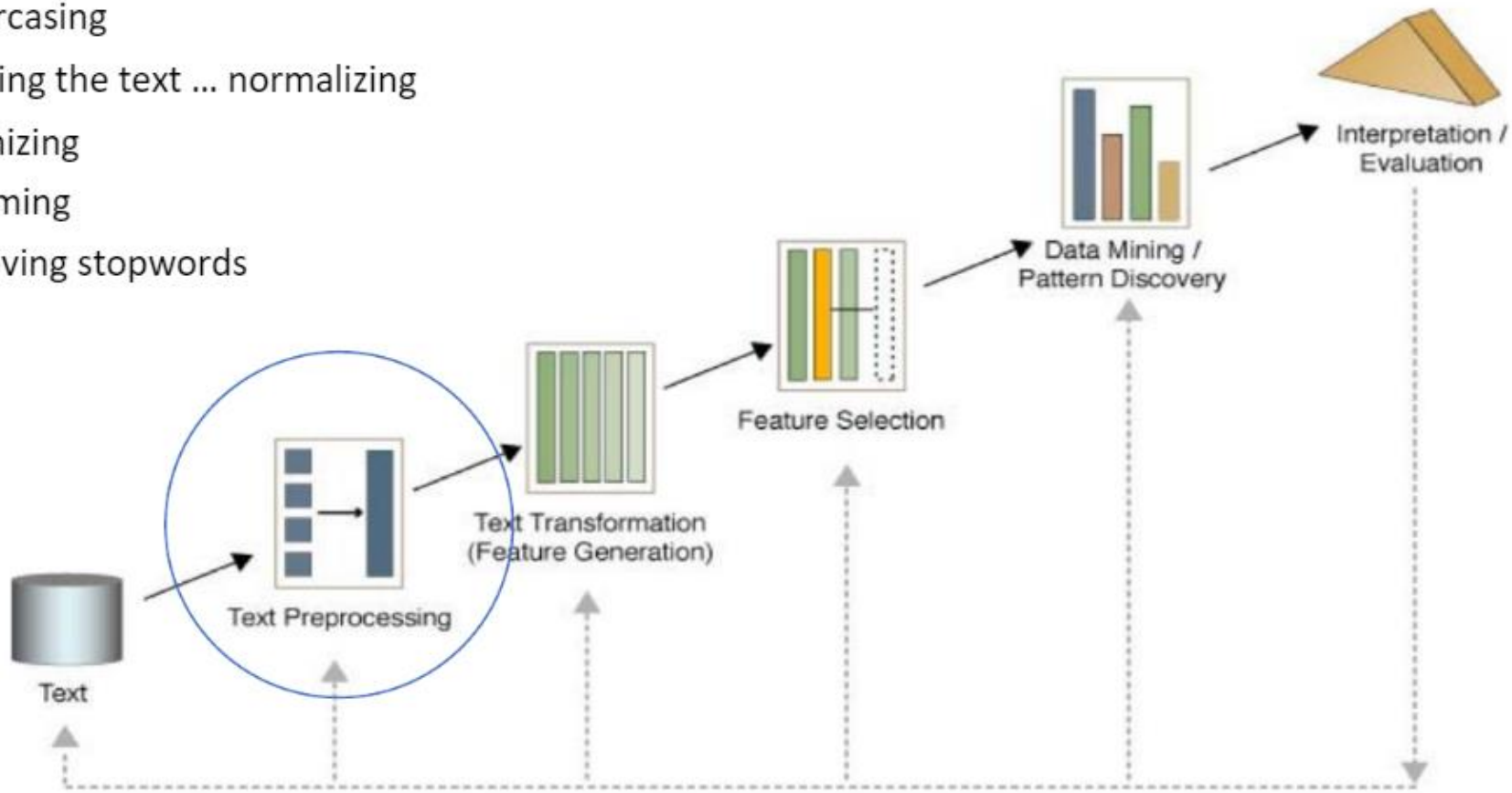
	Tokens = N	**|V|**
Switchboard phone conversations	2.4 million	20 thousand
Shakespeare	884,000	31 thousand
Google N-grams	1 trillion	13 million

Text Preprocessing

TextPreprocessing

Text preprocessing is the task of transforming the text into a form that isanalyzablefor a given task.

- Lowercasing
- Cleaning the text … normalizing
- Tokenizing
- Stemming
- Removing stopwords



Textpreprocessing: lowercasing

- **Lowercasing** all text, although commonly overlooked, is one of the simplest and most effective form of text preprocessing.
- It is applicable to most text mining problems and significantly helps with consistency of expected output.

- This is so that words like Skype and SKYPE are counted as the same thing. Case variations are so common (consider iPhone, iphone, and IPHONE) that case normalization is usually necessary.

- Example of a task where lowercasing is not helpful:
  - Distinguishing US and us,
  - Predicting programming language of a source code file.
  - The word System in Java is quite different from system in python. Lowercasing the two makes them identical, causing the classifier to lose important predictive features.

Textpreprocessing: Cleaning the text & normalizing

Removing HTML tags

- If the reviews or texts are web scraped, chances are they will contain some HTML tags. Since these tags are not useful for text mining tasks, it is better to remove them.

Converting accented characters to ASCII characters

- Words with accents like "latté" and "café" can be converted and standardized to just "latte" and "cafe", else a model will treat them as different words even though they are referring to the same thing.

Expanding contractions

- Contractions are shortened words, e.g., don't and can't. Expanding such words to "do not" and "can not” helps to standardize text.

Standardizing different spelling ... abbreviations

- This is especially important for noisy texts such as social media comments, text messages and comments to blog posts where abbreviations and misspellings are common (2morrow and tomorrow).

Removing extra whitespaces

Removing punctation and special characters

- (matching USA and U.S.A.)

Converting number words to numeric form, removing numbers

Preprocessing: Stop Words

Hi Mr. Smith! I'm going to buy some vegetables (tomatoes and cucumbers) from the store. Should I pick up some black-eyed peas as well?

What is the most frequent term in the text above? Is that information meaningful?

Stop words are words that have very little semantic value.

There are language and context-specific stop word lists online that you can use.

Textpreprocessing: tokenization

- **Tokenization** is a step which splits longer strings of text into smaller pieces, or tokens.
- Larger chunks of text can be tokenized into sentences, sentences can be tokenized into words, etc.
- Tokenization is also referred to as text segmentation or lexical analysis.
- Sometimes segmentation is used to refer to the breakdown of a large chunk of text into pieces larger than words (e.g. paragraphs or sentences), while tokenization is reserved for the breakdown process which results exclusively in words.

Textpreprocessing: tokenization

- How are sentences identified within larger bodies of text?

- Using "sentence-ending punctuation," is ambiguous.

  The quick brown fox jumps over the lazy dog.

- But what about this one:

  Dr. Ford did not ask Col. Mustard the name of Mr. Smith's dog.

- Or this one:

  "What is all the fuss about?" asked Mr. Peters.

- And that's just sentences. What about words? Easy, Right?

  This full-time student isn't living in on-campus housing, and she's not wanting to visit Karachi.

Textpreprocessing: tokenization in different language

- German noun compounds are not segmented
  - ***Lebensversicherungsgesellschaftsangestellter***
  - 'life insurance company employee'
  - German information retrieval needs **compound splitter**

- Chinese and Japanese no spaces between words:
  - 莎拉波娃现在居住在美国东南部的佛罗里达。
  - 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
  - Sharapova now lives in US southeastern Florida
- Further complicated in Japanese, with multiple alphabets intermingled
  - Dates/amounts in multiple formats

フォーチュン500社は情報不足のため時間あた$500K(約6,000万円)

Katakana Hiragana Kanji Romaji

End-user can express query entirely in hiragana!

French

- ***L'ensemble*** → one token or two?
  - ***L*** ? ***L'*** ? ***Le*** ?
  - Want ***l'ensemble*** to match with ***un ensemble***

- Also called **Word Segmentation**
- Chinese words are composed of characters
  - Characters are generally 1 syllable and 1 morpheme.
  - Average word is 2.4 characters long.
- Standard baseline segmentation algorithm:
  - Maximum Matching (also called Greedy)

Issues in Tokenization

- *Finland's capital* → *Finland Finlands Finland's ?*
- *what're, I'm, isn't* → *What are, I am, is not*
- *Hewlett-Packard* → *Hewlett Packard ?*
- *state-of-the-art* → *state of the art ?*
- *Lowercase* → *lower-case lowercase lower case ?*
- *San Francisco* → *one token or two?*
- *m.p.h., PhD.* → *??*

In Natural Language Processing we care about punctuation – and do not discard it!!!!

Finland's capital → *Finland ' s capital*

Tokenization: N-grams

- The most common tokenization process is whitespace/ unigram tokenization.
- In this process entire text is split into words by splitting them from whitespaces.
- However, in some models where every individual word is a token (term), word order is discarded.
- And, in some cases, word order is important, and we want to preserve some information about it in the representation.

- A next step up in complexity is to include sequences of adjacent words as terms.
  - For example, we could include pairs of adjacent words so that if a document contained the sentence The quick brown fox jumps." it would be transformed into
    - set of its constituent words {quick, brown, fox, jumps}
    - Tokens quick_brown, brown_fox, and fox_jumps
- Adjacent pairs are commonly called bi-grams.
- This general representation tactic is called n-grams.

Tokenization: N-grams

- N-grams are useful when particular phrases are significant, but their component words may not be.

- N-grams **advantage** is that they are easy to generate, and they require no linguistic knowledge or complex parsing algorithm.
- The main **disadvantage** of n-grams is that they greatly increase the size of the feature set.
- There are many adjacent word pairs, and still more adjacent word triples.
- The number of features generated can quickly get out of hand, and many of them will be very rare, occurring only once in the corpus.
- Data mining using n-grams almost always needs some special consideration for dealing with massive numbers of features, such as a feature selection stage or special consideration to computational storage space.

Tokenization: Regular Expressions

Let's say you want to tokenize by some other type of grouping or pattern.

Regular expressions (regex) allows you to do so.

Some examples of regular expressions:

- Find white spaces: \s+
- Find words starting with capital letters: [A-Z]['\w]+

- `[A-Z]`: matches any capital letter from A to Z.
- `['\w']`: matches any word character, which includes letters, digits, and underscores.

Code: Tokenization (Regular Expressions)

Input:

```
from nltk.tokenize import RegexpTokenizer

# RegexpTokenizer to match only capitalized words
cap_tokenizer = RegexpTokenizer("[A-Z]['\w]+")
print(cap_tokenizer.tokenize(my_text))
```

Output:

```
['Hi', 'Mr', 'Smith', 'Should']
```

Textpreprocessing: Stemming

- **Stemming** is the process of reducing inflected words to their **word stem**
- Inflection refers to a process ofwordformation in which items are added to the base form of awordto express grammaticalmeanings. The inflection *-ed* is often used to indicate the past tense, changing *walk* to *walked* and *listen* to *listened*.
- Stemming is the process of eliminating affixes (suffixed, prefixes, infixes, circumfixes) from a word in order to obtain a word stem.
- It is language dependent. In some languages, it is more useful than in others (for some tasks German, Spanish and Finnish better performance than in English)

- e.g., ***automate(s), automatic, automation*** all reduced to ***automat***.

for example compressed and compression are both accepted as equivalent to compress.

→

for exampl compress and compress ar both accept as equival to compress

Porter'salgorithm: themostcommonEnglishstemmer

Step 1a

sses → ss	caresses → caress
ies → i	ponies → poni
ss → ss	caress → caress
s → ø	cats → cat

Step 1b

(*v*)ing → ø	walking	→ walk
	sing	→ sing
(*v*)ed → ø	plastered	→ plaster

…

Step 2 (for long stems)

ational→ ate	relational	→ relate
izer→ ize	digitizer	→ digitize
ator→ ate	operator	→ operate

…

Step 3 (for longer stems)

al → ø	revival	→ reviv
able → ø	adjustable	→ adjust
ate → ø	activate	→ activ

…

Textpreprocessing: Lemmization

- **Lemmatization** is the process of converting a word to its base form (correct dictionary headword).
- The output of lemmatization is a root word called a **lemma**. For example
  - *am, are, is* will be converted to *be*
  - *running, runs, ran* will be replaced by *run*
- Stemming and Lemmatization both generate the foundation of the inflected words with the difference being that stem may not be an actual word whereas, lemma is an actual language word.
- Stemming follows an algorithm with steps to perform on the words which makes it faster. Whereas, in lemmatization, we need a corpus also to supply lemma which makes it slower than stemming. Furthermore, we might need to define a parts-of-speech (noun,pronoun,verb, adjective, adverb,preposition, conjunction, and interjection) to get the proper lemma (to distinguish *running* verb vs *running* noun)

NLP Toolkits

- NLTK (Natural Language Toolkit)
  - The most popular NLP library
- TextBlob
  - Wraps around NLTK and makes it easier to use
- spaCy
  - Built on Cython, so it's fast and powerful
- gensim
  - Great for topic modeling and document similarity

THANK YOU