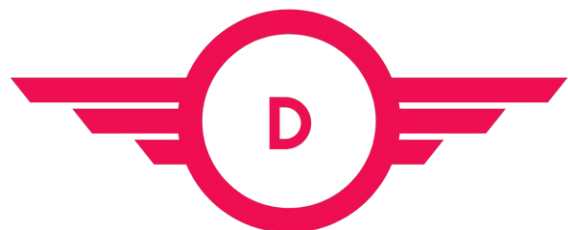


COS 760 - Natural Language Processing

Lecture 1 - 12/02/2025

Dr Abiodun Modupe
abiodun.modupe@cs.up.ac.za
Prof Vukosi Marivate,
ABSA UP Chair of Data Science
vukosi.marivate@cs.up.ac.za



Data Science for Social Impact



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za



Course Lectures

Dr Abiodun Modupe

- Senior Member, Data Science for Social Impact
- Artificial Intelligence
- Cybersecurity
- Computer Vision
- Natural Language Processing

Telephone: 012 420 5232

Email: abiodun.modupe@cs.up.ac.za



**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Course Lectures

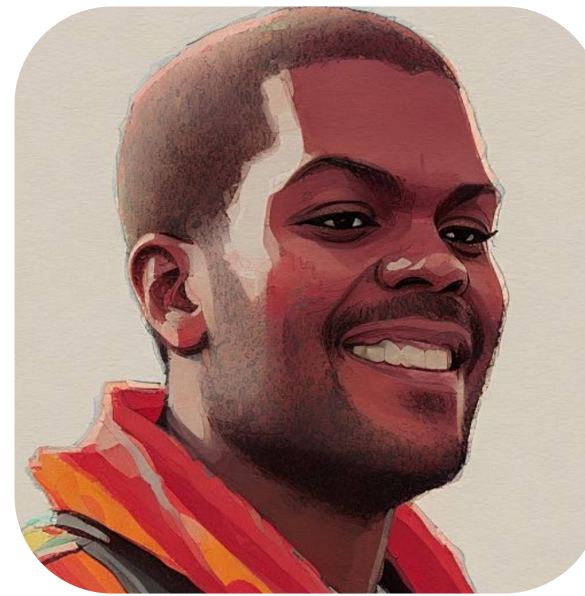
Prof Vukosi Marivate

ABSA Data Science Chair

- Data Science for Social Impact
- Machine Learning
- Natural Language Processing
- Anomaly Detection

Telephone: 012 420 3561

Email: vukosi.marivate@cs.up.ac.za



**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Credits - 15

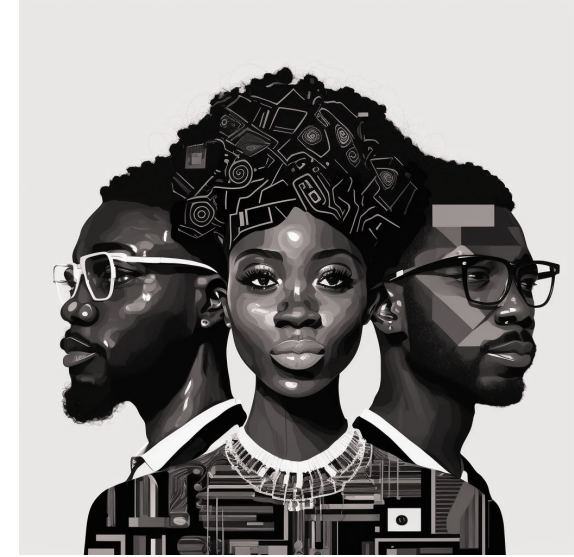
15 Credits = 150 Hours on working on this class

15 Lectures = approx 30 hours

Assignments/Readings = approx 60 hours

Project = approx 50 hours

Other activities = approx 10 hours



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Assessment for this class

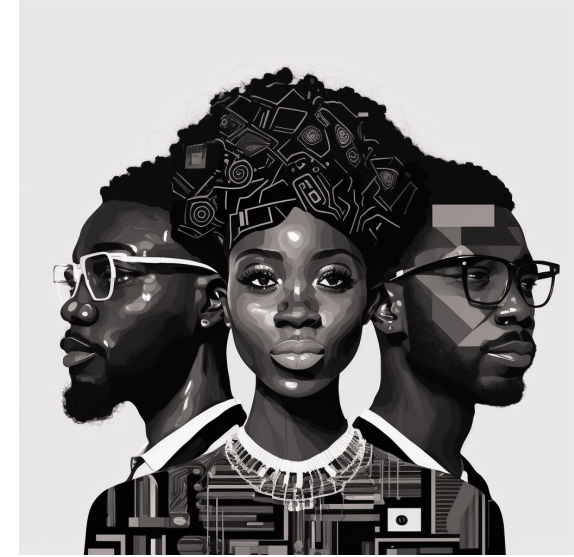
Assignment [50%]

1. HW 1. TFIDF, Simple Classifiers [10%]
2. HW 2. Topic Modelling, Clustering + Embeddings [10%]
3. HW 3. Neural Network, BERT and Use case [15%]
4. HW 4. LLM [15%]

Project [50%]

1. Proposal [10 %]
2. Presentation [10 %]
3. Poster [10 %]
4. Report [20 %]

The student must obtain a final mark of at least 50% to pass the module.



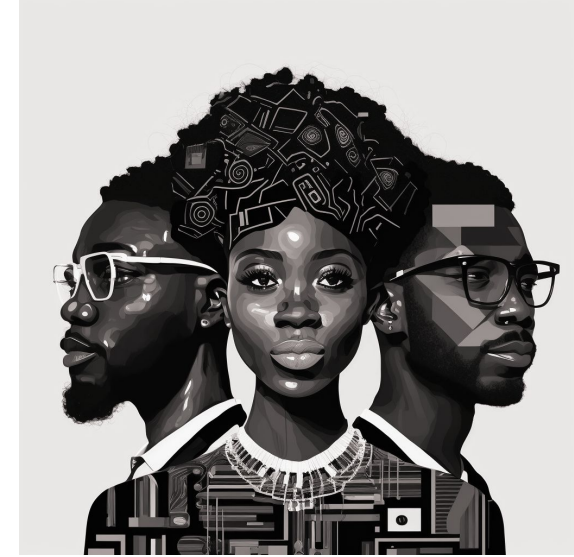
**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Suggested Reading

Primary

Speech and Language Processing [SLP]
(Third Edition draft: 2025 pre-release)
by: Dan Jurafsky and James H. Martin.



Secondary

Natural Language Understanding with Distributed Representation
[NLUDR]
By: Kyunghyun Cho 2015.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknoloṽi ya Tshedimošo

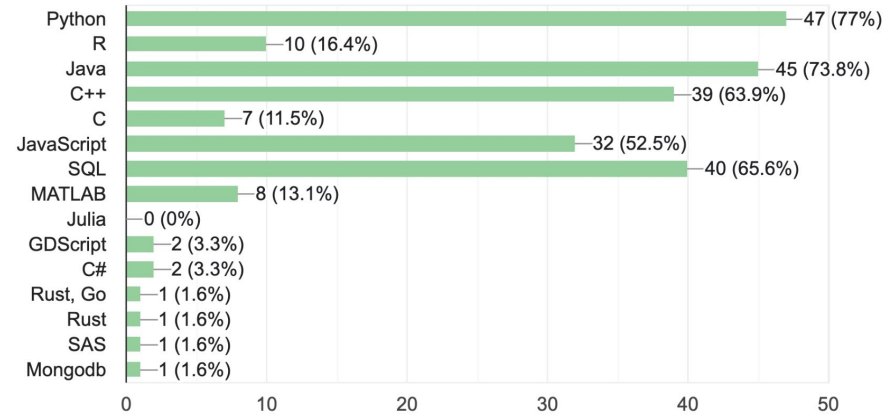
Make today matter

www.up.ac.za

Survey

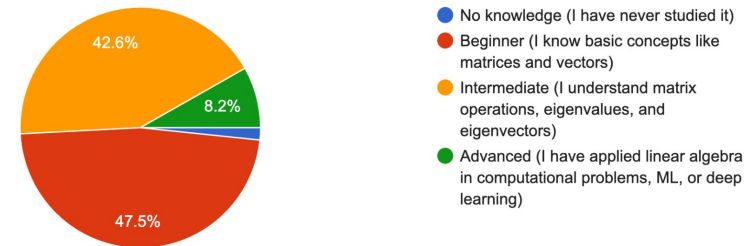
Which programming languages are you comfortable with?

61 responses



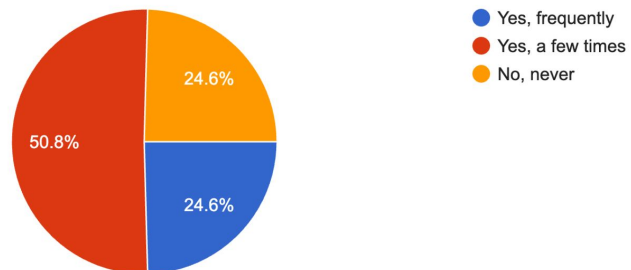
How would you rate your knowledge of linear algebra?

61 responses



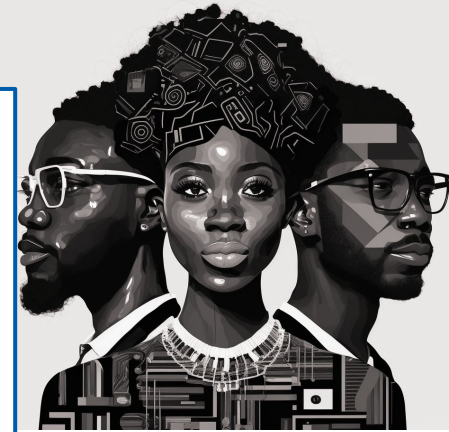
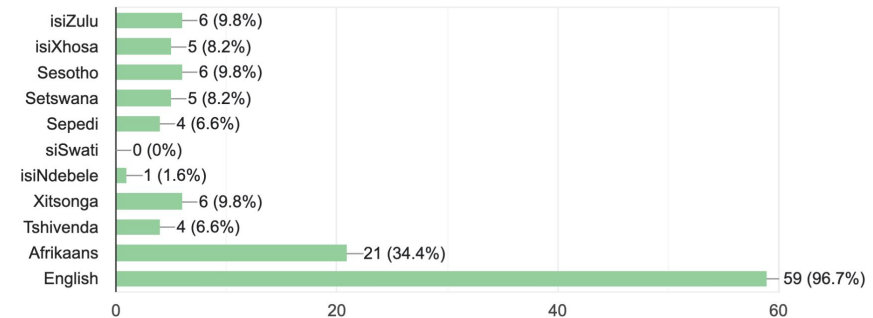
Have you used Jupyter Notebooks before?

61 responses



Which South African languages can you read fluently?

61 responses



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

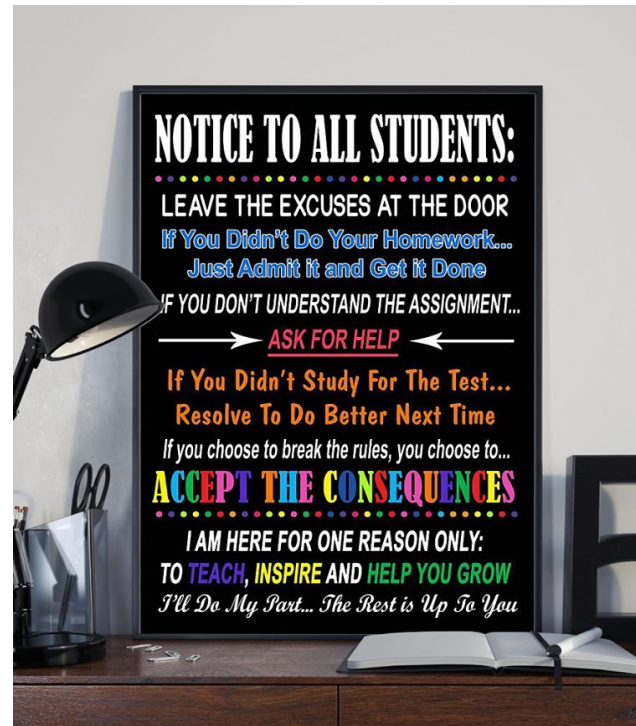
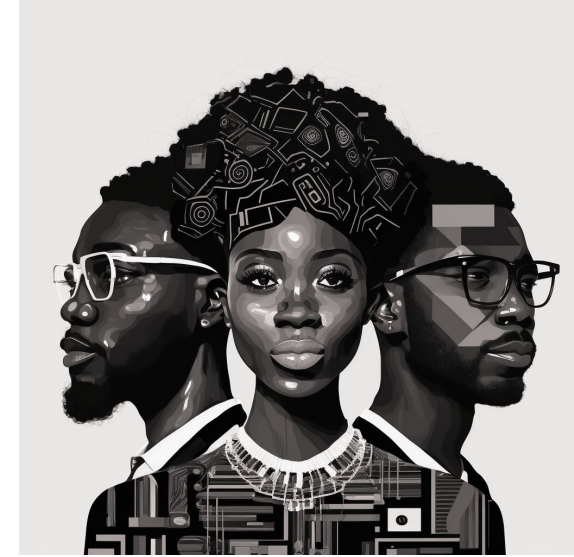
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Before we start

- Be generous [Share yourself and your knowledge]
- Ask Questions [Even on the discussion group]
- Don't pretend to know stuff [Enquiry leads to breakthroughs]



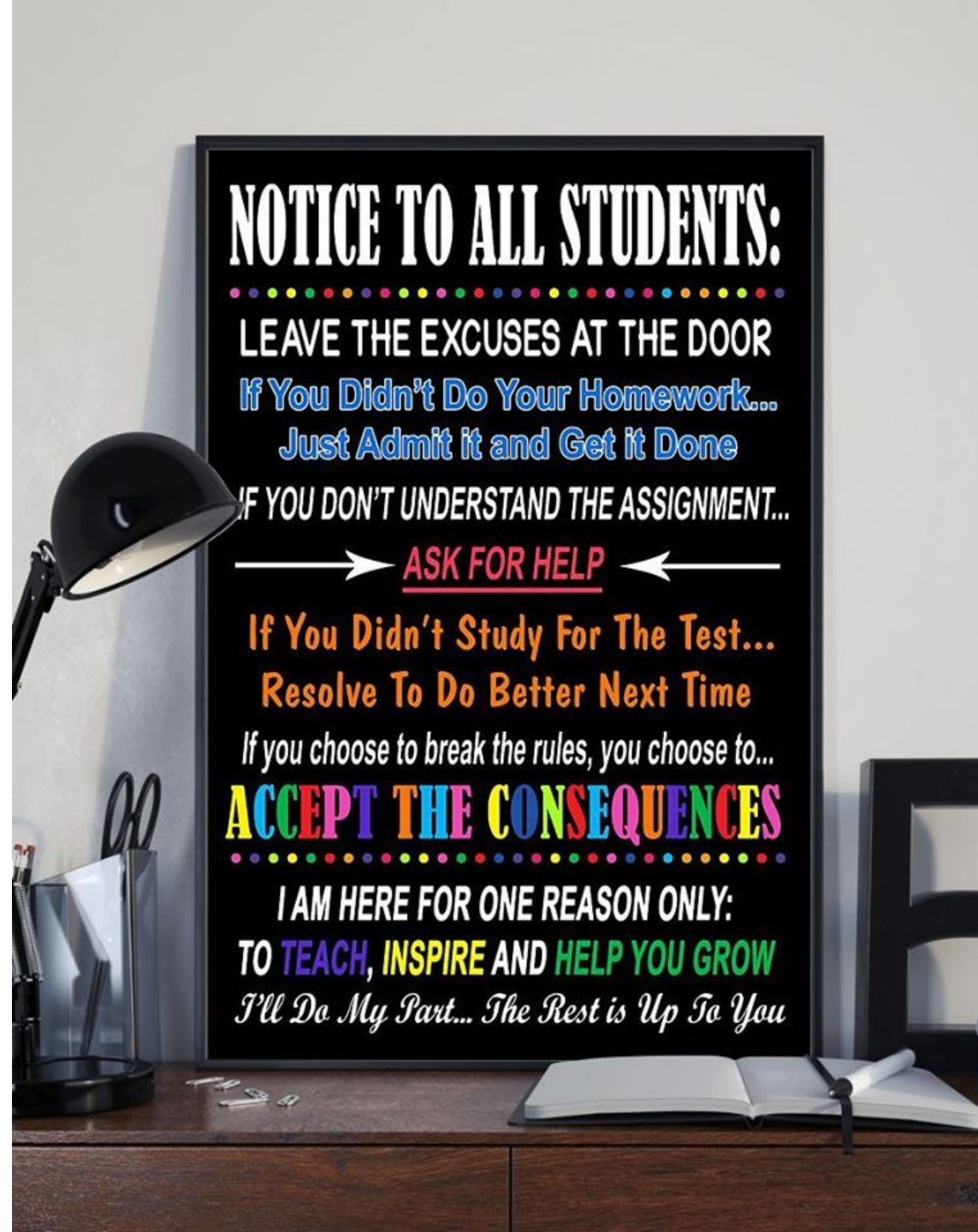
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknoloṱši ya Tshedimošo

Make today matter

www.up.ac.za



Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Introduction to Natural Language Processing



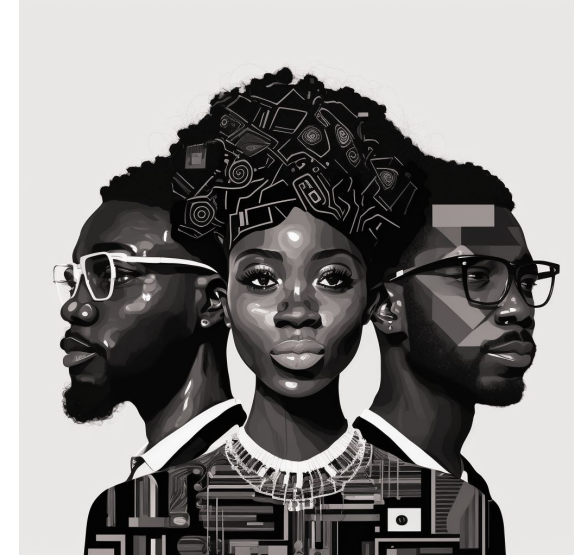
Outline of today (1 hour)

Understanding

- What is NLP
 - Historical basis
 - Evolution
 - Engineering
 - Science

Through

- Traditional NLP [This lecture] (Chap 1-2 of [SLP])



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

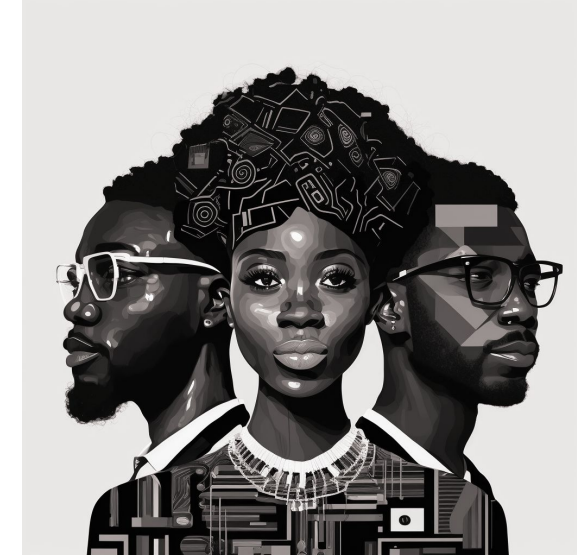
**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

What is Natural Language Processing (NLP)?



Definition of NLP: A field at the intersection of linguistics and computer science that focuses on making human language **computable**.

Importance of NLP in modern society:

- Information retrieval (Google Search, Wikipedia)
- Communication tools (translation, chatbots)
- Text analytics (news classification, sentiment analysis)
- Speech applications (virtual assistants, transcription)

Historical Perspective:

- Early symbolic and rule-based approaches (before statistical methods took over).
- The evolution of NLP from linguistics to computational models.



**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Fundamental Challenges in NLP

Ambiguity in Language:

- **Lexical ambiguity** ("bank" → financial institution vs. riverbank)
- **Syntactic ambiguity** ("I saw the man with the telescope.")
- **Semantic ambiguity** ("The chicken is ready to eat" → Who eats whom?)

Morphological Complexity:

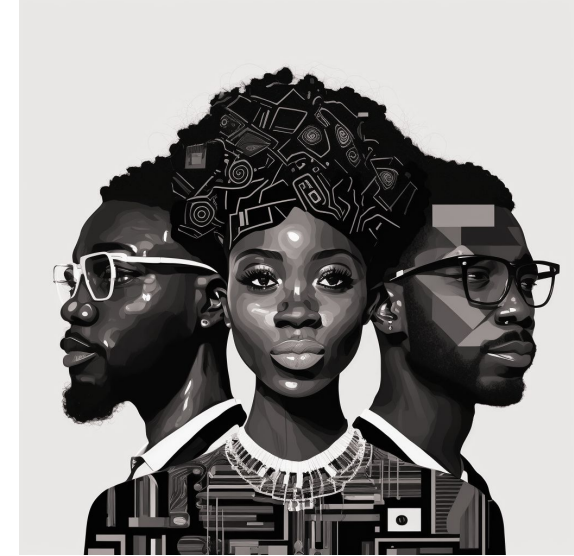
- Word formation (root words, inflections, compound words)
- Differences in morphology across languages (e.g., agglutinative vs. isolating languages)

Context Dependence:

- Words change meaning based on **context** (e.g., "hot" → weather vs. food vs. performance)
- The role of **pragmatics** in NLP (e.g., sarcasm, politeness, indirect speech)

Low-Resource Languages:

- The problem of **data scarcity** for African and indigenous languages.
- Existing efforts: Masakhane, AfriBERTa, Lelapa AI, etc.



Faculty of Engineering,
Built Environment and
Information Technology

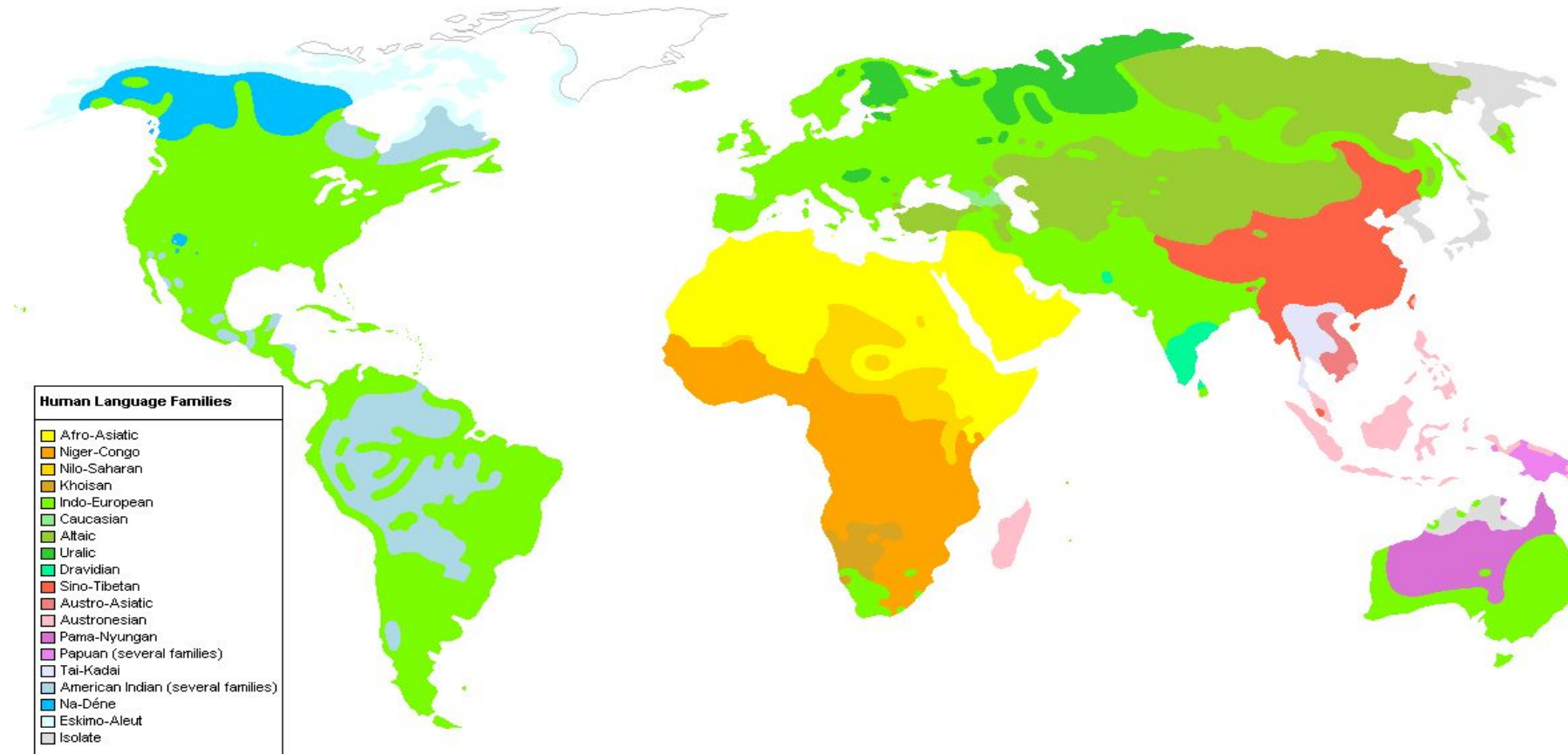
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Natural Language Processing



Why NLP?

- Understanding our world
- Communication
- Spreading of ideas



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Theoretical Basis of NLP – Chomsky's Linguistic Theories & Generative Grammar

What is Language?

- A structured system of communication with syntax, semantics, and pragmatics.
- Theoretical perspectives on language shape how NLP systems are designed.

Chomsky's Linguistic Theories

1. *Universal Grammar (UG)*

- Humans are born with an innate ability to learn language.
- All languages share common underlying structures.
- Implication for NLP: Can we build systems that mimic this innate structure?

2. *Chomsky Hierarchy of Languages (Formal Language Theory)*

- Regular Languages → Finite State Automata (FSA) → Simple tokenization, regex.
- Context-Free Languages (CFLs) → Pushdown Automata → Syntax parsing.
- Context-Sensitive Languages → More complex grammars, requiring deep structures.
- Turing-Complete Languages → Theoretically infinite computational power.



Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknološi ya Tshedimošo

Theoretical Basis of NLP – Chomsky's Linguistic Theories & Generative Grammar

Generative Grammar & NLP

- Phrase Structure Grammar: Sentences are formed using a set of rules.
- Transformational Grammar: Language structure is generated via transformations.
- Dependency Grammar: Emphasizes relationships between words (used in modern NLP parsing).



NLP Implication:

- Rule-based NLP systems (early machine translation) relied on explicit grammar rules.
- **Limitations:** Hard to scale across languages, too rigid for real-world usage.



Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

The Evolution of NLP – From Rules to Data-Driven Methods

Early Approaches: Rule-Based & Symbolic NLP

- *Expert Systems*: Hand-crafted rules (e.g., rule-based translation, syntax trees).

Challenges:

- Hard to scale.
- Struggles with ambiguity & real-world variation.

Statistical NLP (1990s–2010s)

- Inspired by probability & data-driven linguistics.

Key Techniques:

- n-Gram Models: Predict words based on prior context. (example later)
- Hidden Markov Models (HMMs): Used in speech recognition, POS tagging.
- Probabilistic Context-Free Grammars (PCFGs): Weighted parsing rules.

Limitations:

- Requires large annotated datasets.
- Does not capture deep language understanding.



Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

The Evolution of NLP – From Rules to Data-Driven Methods

Deep Learning & NLP (2010s–Present)

- Uses neural networks to learn patterns from raw text data.

Key Innovations:

- *Word Embeddings (Word2Vec, GloVe)*: Represent words as vectors.
- *Recurrent Neural Networks (RNNs, LSTMs)*: Capture sequence relationships.
- *Transformers (BERT, GPT)*: Context-aware models trained on massive text data.

Advantages:

- Learns implicit rules instead of needing manually defined ones.
- Handles ambiguity better than rule-based methods.

Challenges:

- Requires huge amounts of data & computing power.
- Can still struggle with bias, hallucination, and reasoning.

 **NLP Today: Hybrid models combine linguistic rules + statistical/deep learning to balance efficiency & interpretability.**



Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

NLP – Engineering vs. Cognitive Problem

Dual Nature of NLP

- Engineering → Build efficient language models.
- Cognitive → Understand how humans process language.

1. NLP as Engineering

Goal: Process text at scale.

- Challenges: Ambiguity, data scarcity, scalability.

Methods:

- Rule-Based → Grammar rules.
- Data-Driven → Statistical & deep learning.

2. NLP as Cognition

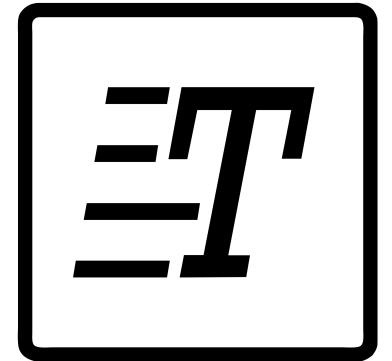
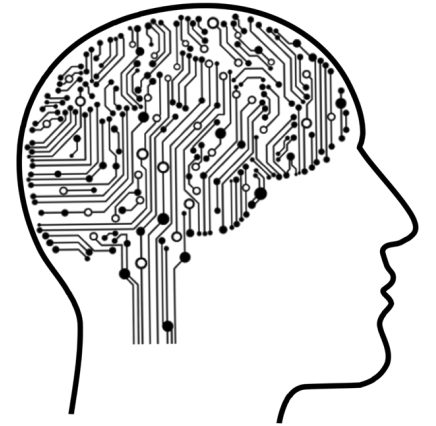
Goal: Model human language understanding.

Key Questions:

- How do humans learn & infer meaning?
- Can NLP models replicate reasoning?

Influences: Chomsky's Universal Grammar, Psycholinguistics.




♦ **Takeaway: NLP needs both engineering & cognition for better AI.**






Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Symbolic vs. Connectionist NLP

1. Symbolic (Rule-Based) NLP

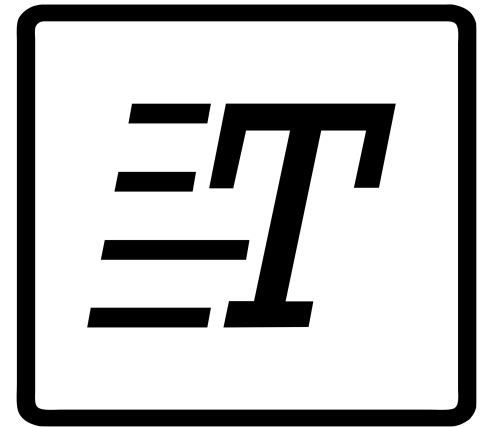
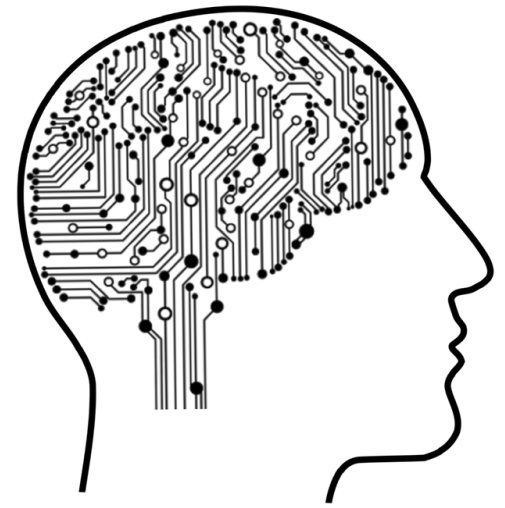
- Linguistics-driven, predefined grammar.
-  Interpretable, structured tasks.
-  Hard to scale, struggles with ambiguity.
-  Example: Rule-based chatbots.

2. Connectionist (Neural) NLP

- Learns patterns from data.
-  Handles context, ambiguity, self-learns.
-  Opaque, needs big data.
-  Example: BERT, GPT.

3. Hybrid NLP

- Combining rules + deep learning for explainability & power.
-  Example: Neuro-Symbolic AI.
- ♦ ***Takeaway: Future NLP blends structure & learning.***

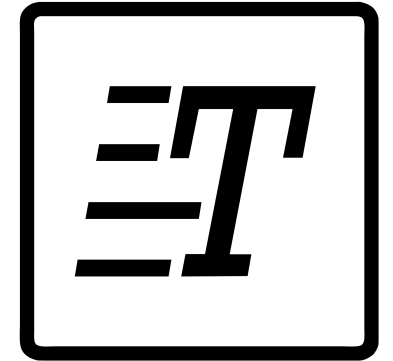
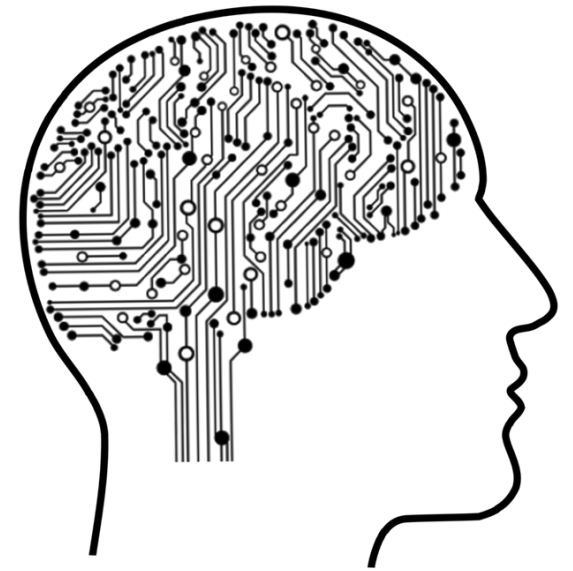


Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknološhi ya Tshedimošo

NLP as tasks

NLP is a broad field, encompassing a variety of tasks, including:

- Part-of-speech (PoS) tagging: noun, verb, adjective, etc.)
- Named entity recognition (NER): person names, organizations, locations, etc.
- Question answering
- Speech recognition
- Text-to-speech (T2S) and Speech-to-text (S2T)
- Topic modeling
- Sentiment classification
- Language modeling
- Translation



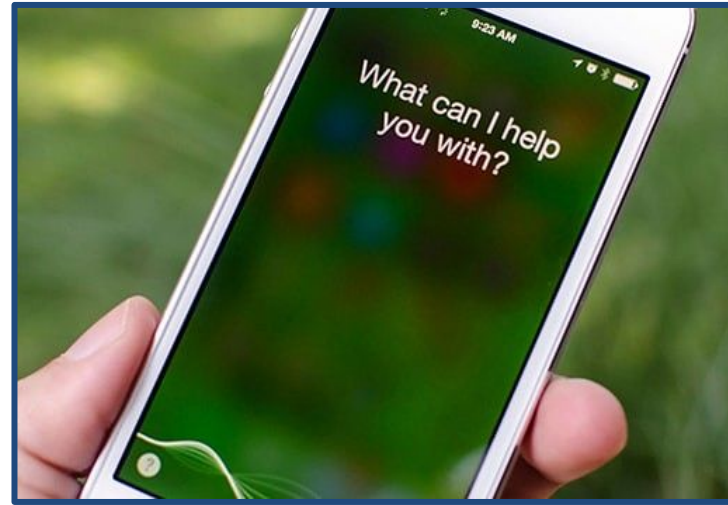
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknološi ya Tshedimošo

NLP Breakthroughs

Virtual Assistants



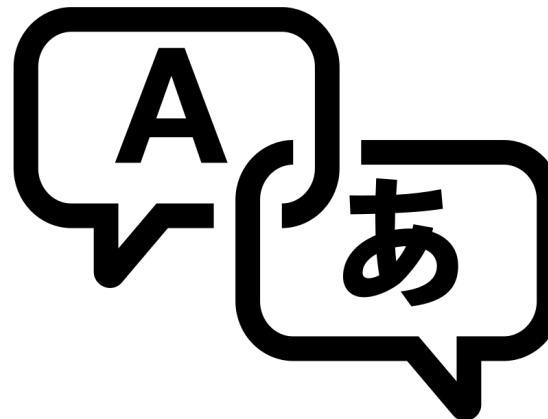
Spam Filters



ChatBots



Translation



NLU?

Natural Language Understanding

Natural-language understanding, or natural-language interpretation, is a subtopic of natural-language processing in artificial intelligence that deals with machine reading comprehension. - Wikipedia



Meaning: *I am thinking.*

Natural Language Processing - Basic Text Processing



Regular Expressions

Looking **for** patterns in text. Example below

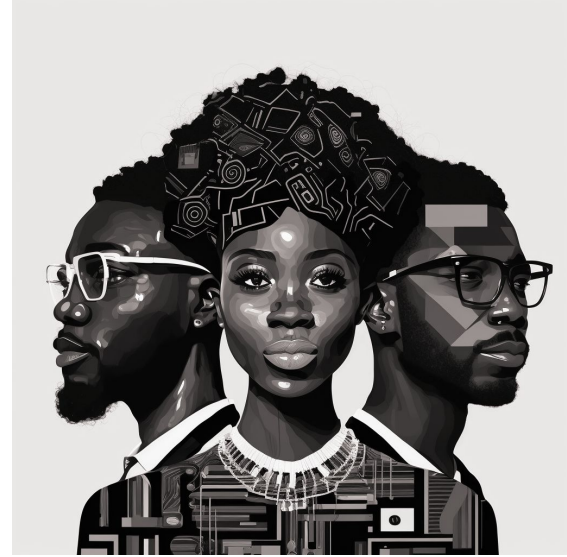
Start of the line

3 to 15 characters long

`^[a-z0-9_-]{3,15}$`

End of the line

letters, numbers, underscores, hyphens

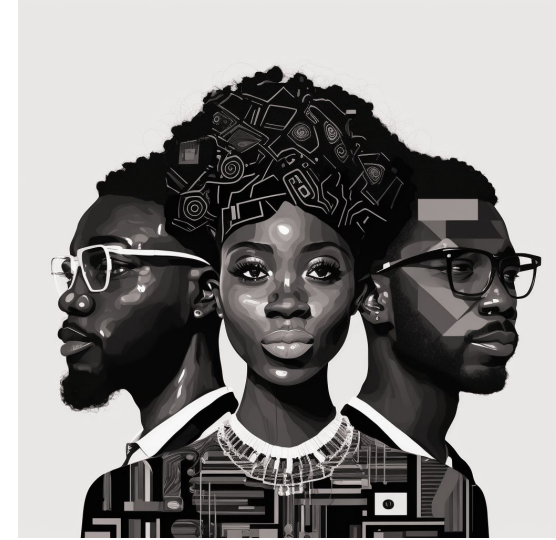


Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

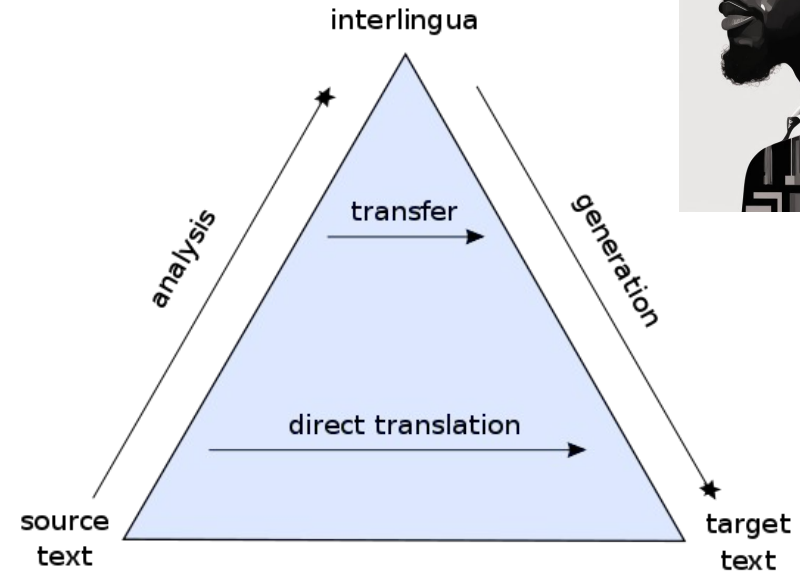
Rule Based Systems

A ↔ あ



Translation: Direct Systems

Translation: Transfer RBMT Systems



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

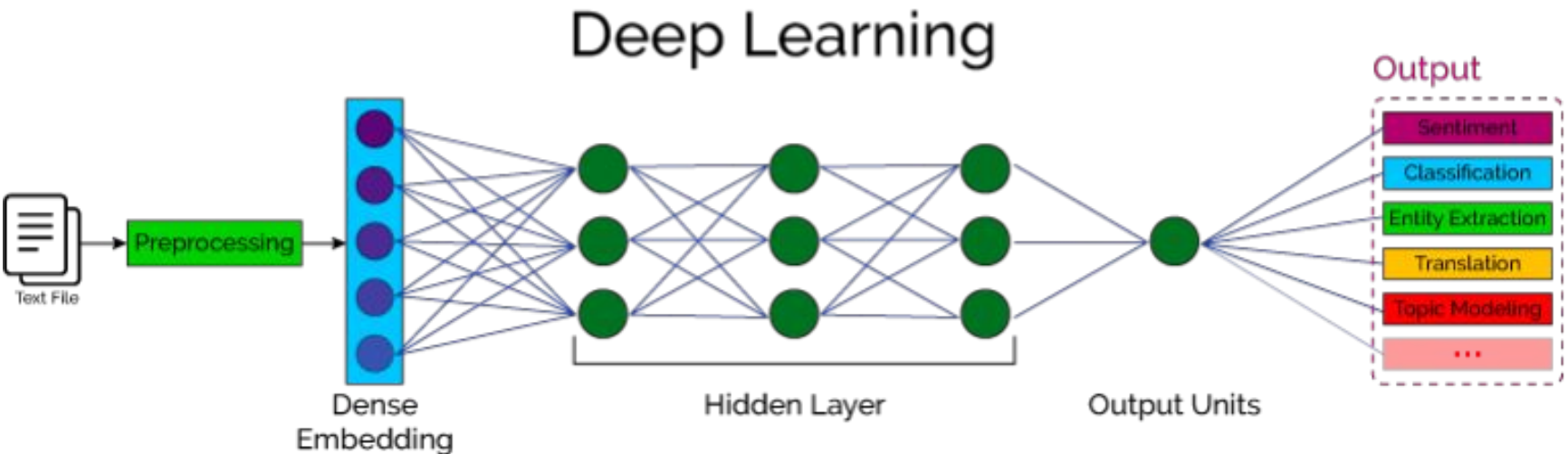
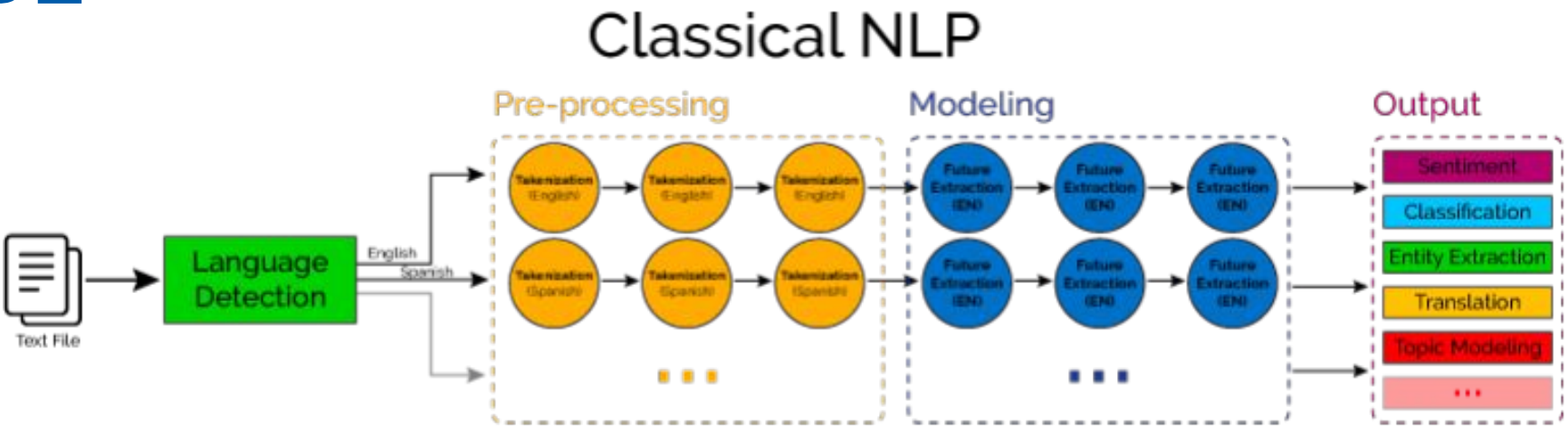
Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Data driven (Statistical) and cutting edge DL



NLP Tasks

Input	Output	Task
Dumelang Bagolo	Greetings, Elders	Translation [1]
It was one of the worst days of the year	Negative	Sentiment Analysis/ Opinion Mining [2]
Who is the president of South Africa?	Cyril Ramaphosa	Question and Answering/Reading Comprehension [3]
You have received a credit of R8,765.50 made to you by Sars eFiling. Please see below reference for confirmation. Sars Online Audit	Suspicious	SPAM Detection
More than 350 elephants have died in northern Botswana in a mysterious mass die-off described by scientists as a “conservation disaster”. A cluster of elephant deaths was first reported in the Okavango Delta in early May, with 169 individuals dead by the end of the month. By mid June, the number had more t....	A cluster of elephant deaths was first reported in the Okavango Delta in early May, with 169 individuals dead by the end of the month.	Summarisation [4]



Faculty of E
Built Enviro
Informatior
Fakulteit Ingenieursw
Inligtingtegnologie / L
Tikologo ya Kago le Th



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

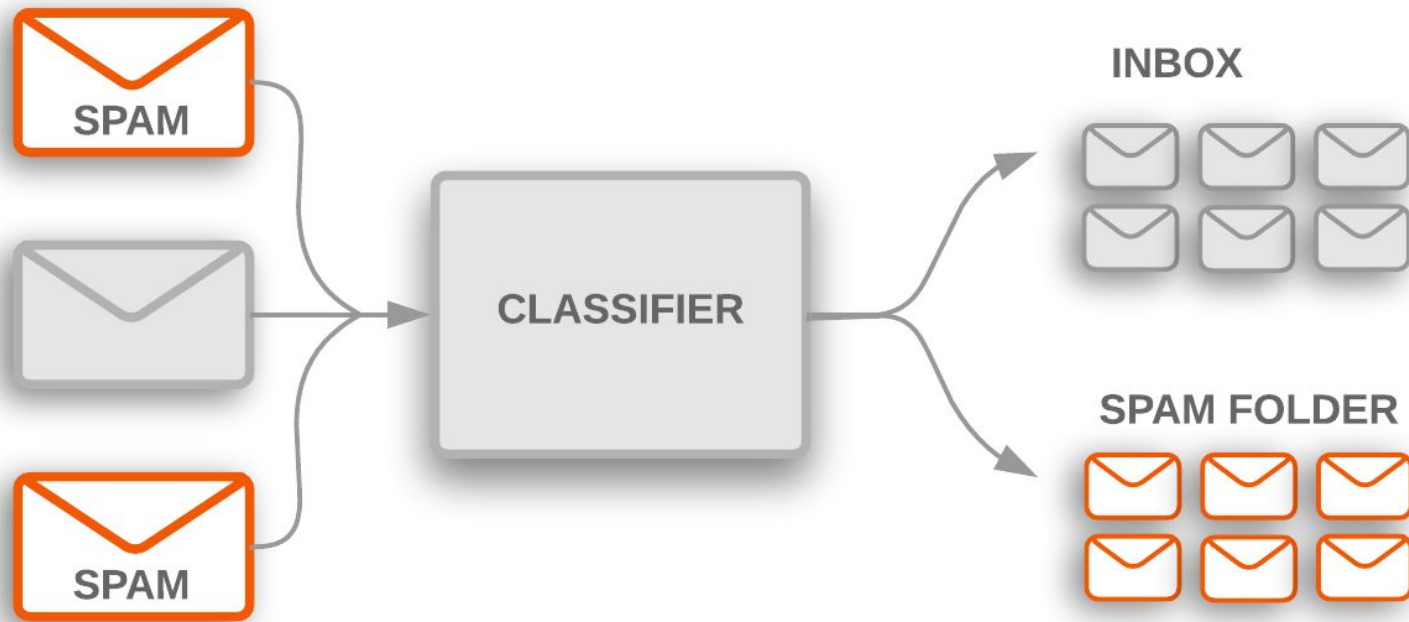
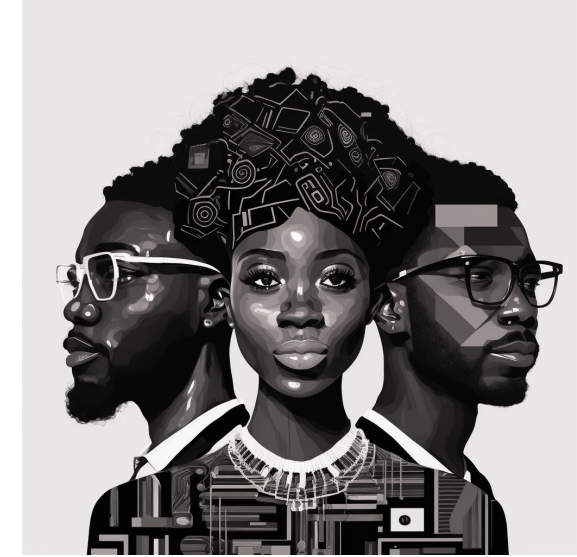
Inspired by G Neubig

- [1] - <https://translate.google.co.za>
- [2] - <https://demo.allennlp.org/sentiment-analysis>
- [3] - <https://demo.allennlp.org/reading-comprehension/>
- [4] - <http://textsummarization.net/text-summarizer>

Make today matter
www.up.ac.za

Make today matter
www.up.ac.za

NLP Example Task 1: Text Classification



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

**Let's develop our
first tool**



Getting Machines to Process Language

We need features!!!

Treat words as **atomic features**

- *car, house, tree, drove, past*

Approach: one hot vector for each symbol

car = [1 0 0 0 0]

house = [0 1 0 0 0]

tree = [0 0 1 0 0]

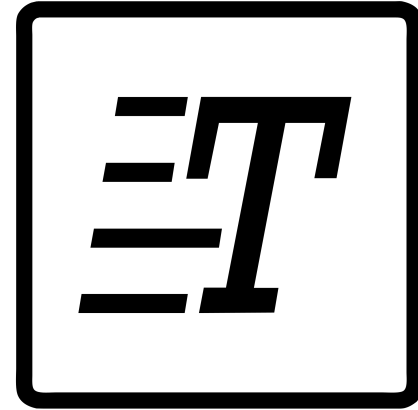
drove = [0 0 0 1 0]

past = [0 0 0 0 1]

Sentences?: We can make sentences/combinations

car drove past tree house = [1 1 1 1 1]

Concatenate -> [1 0 0 0 0] [0 0 0 1 0] [0 0 0 0 1] [0 0 1 0 0] [0 1 0 0 0]



Term Frequency

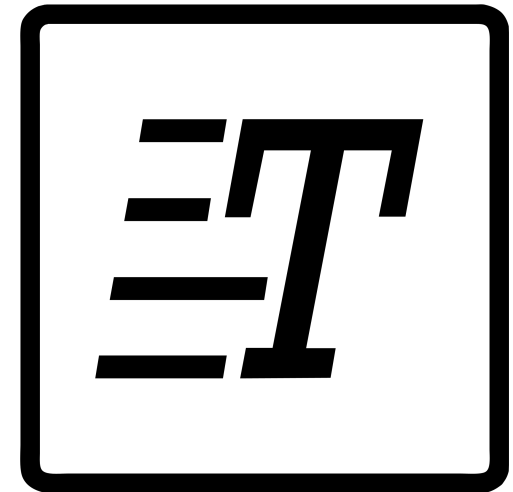
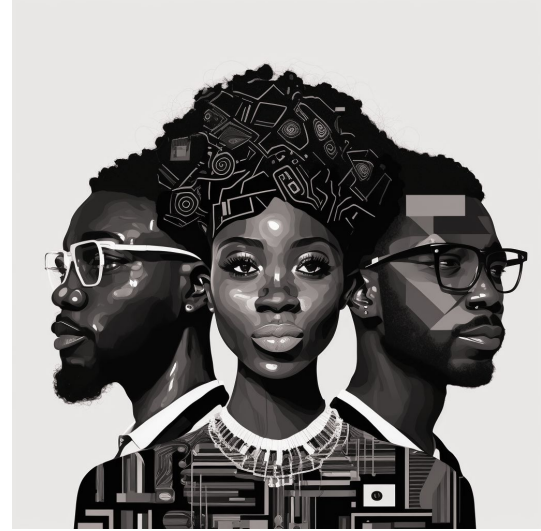
USE CASE Document Retrieval (Simple search Engine)

Task:

- Have many documents (**A corpus**)
- We have a search term **Q** (Q made up of symbols)
- Want to return documents relevant to **Q**

Approach

- Count how many times symbols appear in each document.
- Return the documents that have the highest count of the symbols in **Q**



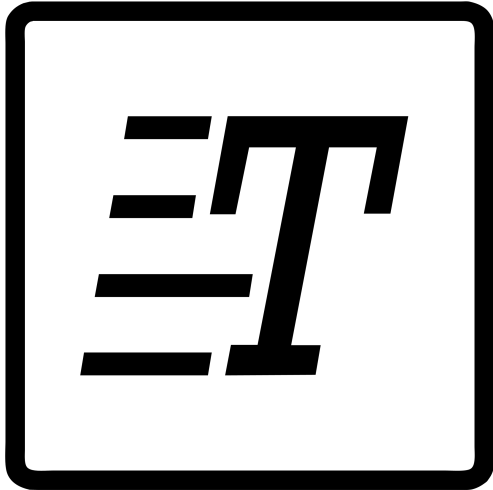
Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknoloṭši ya Tshedimošo

Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...



Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknološi ya Tshedimošo

Tokenization – Breaking Text into Units

Definition: Splitting text into words, subwords, or characters.

Why it Matters: Basis for search, translation, and speech recognition.

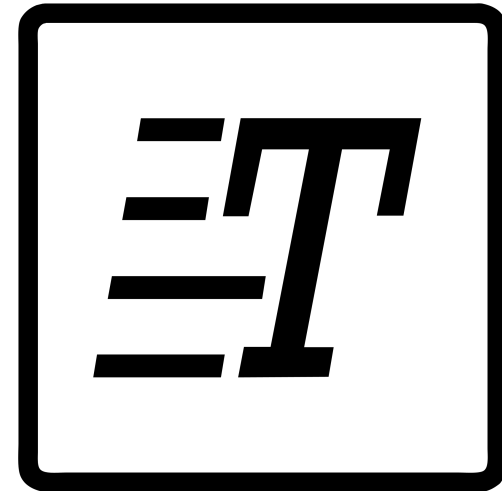
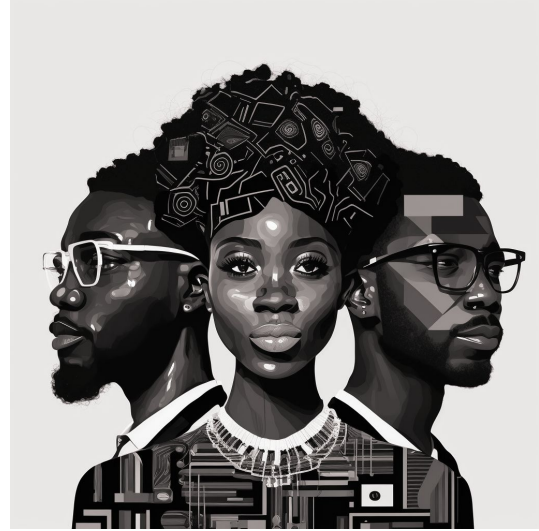
Types:

- Whitespace-based → Splits on spaces (fails for “New York”).
- Subword tokenization → Handles out-of-vocabulary words.
- Byte Pair Encoding (BPE) → Merges frequent character pairs (used in GPT).



Example: "Natural-language processing"

- Whitespace: ["Natural-language", "processing"]
- Subword: ["Natural", "-", "language", "processing"]
- BPE: ["Natur", "al", "-", "lang", "uage", "processing"]



Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknološi ya Tshedimošo

Tokenization

Sentence Tokenization

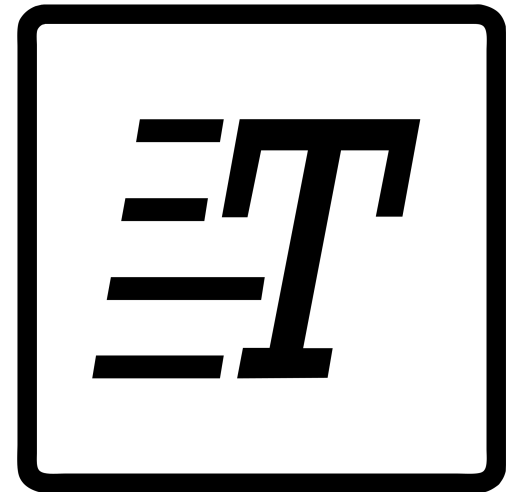
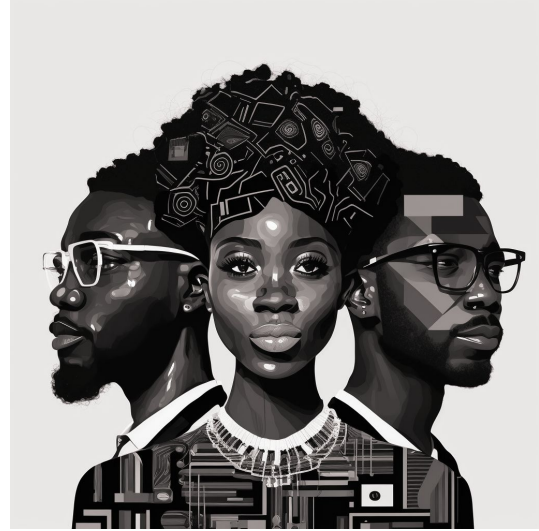
Document

"Wakanda is located in East Africa, although its exact location has varied throughout the nation's publication history: some sources place Wakanda just north of Tanzania, while others—such as Marvel Atlas #2—show it at the north end of Lake Turkana, in between South Sudan, Uganda, Kenya, and Ethiopia (and surrounded by fictional countries like Azania, Canaan, and Narobia). In the Marvel Cinematic Universe (The Black Panther), on-screen maps use the location given in Marvel Atlas #2."

Sentence Tokenization

S1 = Wakanda is located in East Africa, although its exact location has varied throughout the nation's publication history: some sources place Wakanda just north of Tanzania, while others—such as Marvel Atlas #2—show it at the north end of Lake Turkana, in between South Sudan, Uganda, Kenya, and Ethiopia (and surrounded by fictional countries like Azania, Canaan, and Narobia).

S2 = In the Marvel Cinematic Universe (The Black Panther), on-screen maps use the location given in Marvel Atlas #2.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknoloṭši ya Tshedimošo

Tokenization

Word Tokenization

Document

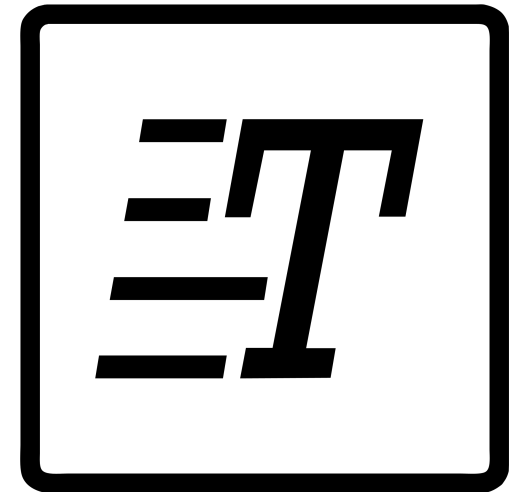
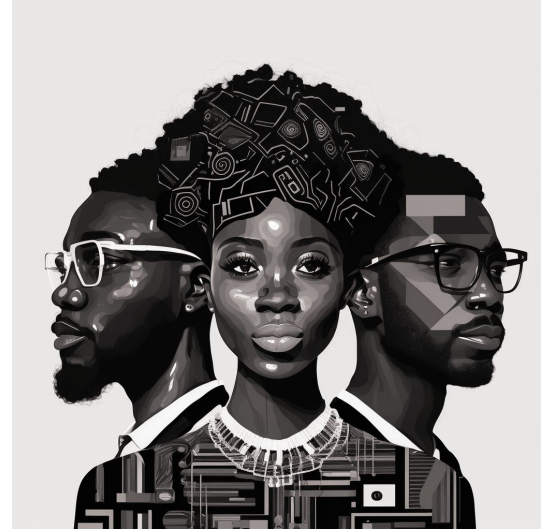
"Wakanda is located in East Africa, although its exact location has varied throughout the nation's publication history: some sources place Wakanda just north of Tanzania, while others—such as Marvel Atlas #2—show it at the north end of Lake Turkana, in between South Sudan, Uganda, Kenya, and Ethiopia (and surrounded by fictional countries like Azania, Canaan, and Narobia). In the Marvel Cinematic Universe (The Black Panther), on-screen maps use the location given in Marvel Atlas #2."

Word Tokenization

S1 = ['Wakanda', 'is', 'located', 'in', 'East', 'Africa', ',', 'although', 'its', 'exact', 'location', 'has', 'varied', '...']

S2 = ['In', 'the', 'Marvel', 'Cinematic', 'Universe', '(', 'The', 'Black', 'Panther', ')', ',', '...']

We could also do character-level extraction!!!!!!



Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Stemming – Reducing Words to Their Root

Definition: Truncates words to their base form.

Example: "running" → "run", "happily" → "happi".

Algorithms:

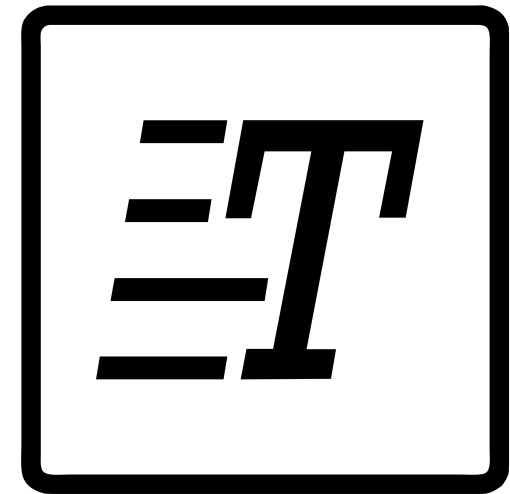
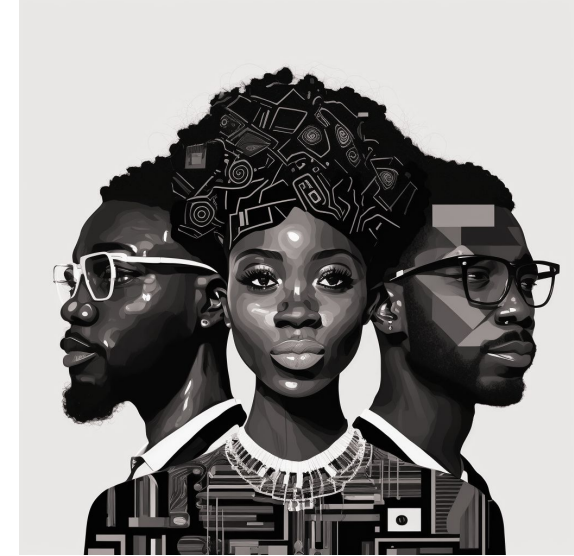
- Porter Stemmer → Rule-based suffix stripping.
- Lancaster Stemmer → More aggressive truncation.

Limitations:

- Not always linguistically accurate.
- "better" → "bet" (incorrect meaning).

Python Example:

```
from nltk.stem import PorterStemmer
ps = PorterStemmer()
print(ps.stem("running"))    # Output: run
```



**Faculty of Engineering,
Built Environment and
Information Technology**
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Lemmatization – More Accurate Word Reduction

Definition: Converts words to their **dictionary base form**.

Example: "running" → "run", "better" → "good".

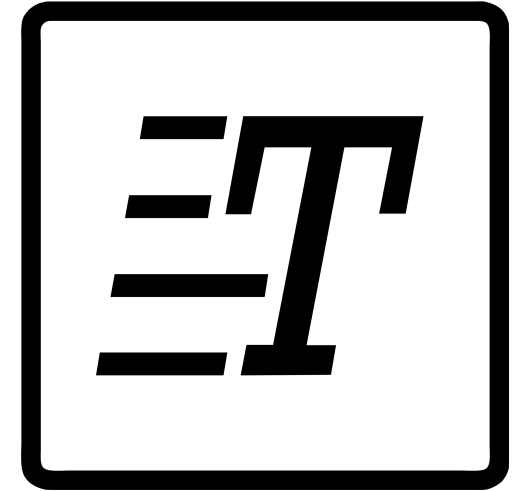
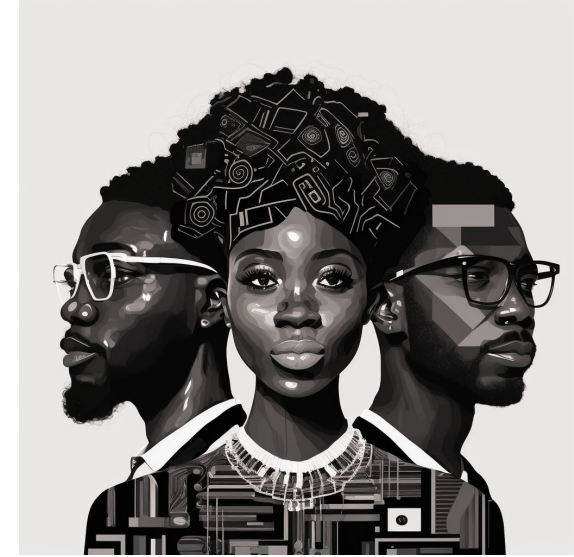
Uses a Lexicon (WordNet) instead of simple rules.

More accurate than stemming (retains meaning).

 **Python Example:**

```
from nltk.stem import WordNetLemmatizer
```

```
lemmatizer = WordNetLemmatizer()  
print(lemmatizer.lemmatize("better", pos="a"))  
# Output: good
```



**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknološi ya Tshedimošo

n-grams

Features do not need to be single words

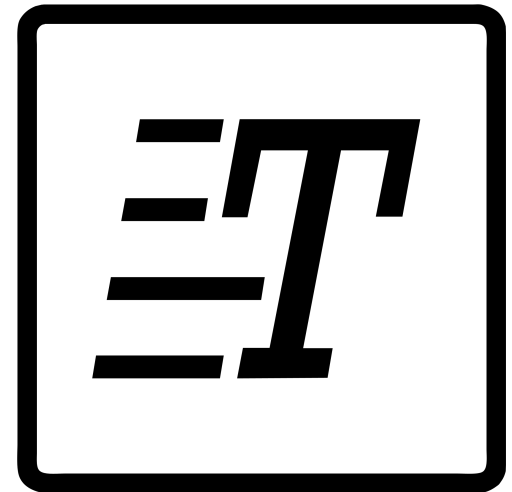
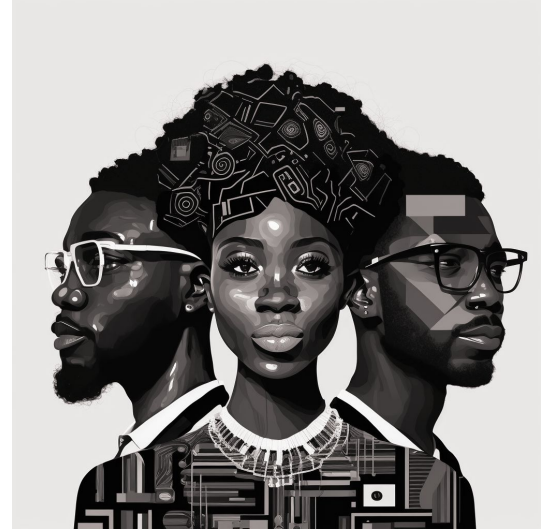
Bi-Grams [pairs of words]

- **Today was** a beautiful day
- Today **was a** beautiful day
- Today was **a beautiful** day

Tri-Grams [pairs of words]

- **Today was a** beautiful day
- Today **was a beautiful** day
- Today was **a beautiful day**

Etc.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

TF-IDF

TF-IDF

TF-IDF is a measure of originality of a word by comparing the number of times a word appears in a doc with the number of docs the word appears in.

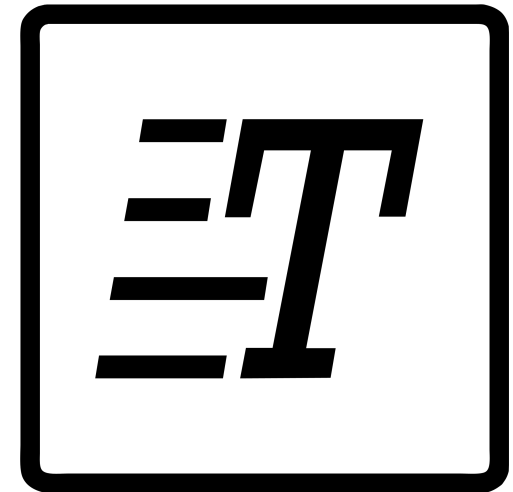
$$\text{TF-IDF} = \text{TF}(t, d) \times \text{IDF}(t)$$

Term frequency

Inverse document frequency

Number of times term t appears in a doc, d

$$\log \frac{1 + \overset{\text{\# of documents}}{n}}{1 + \underset{\text{Document frequency of the term } t}{df(d, t)}}$$



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Regular Expressions (Regex) in NLP

Definition: Pattern matching for text processing.

Common Uses:

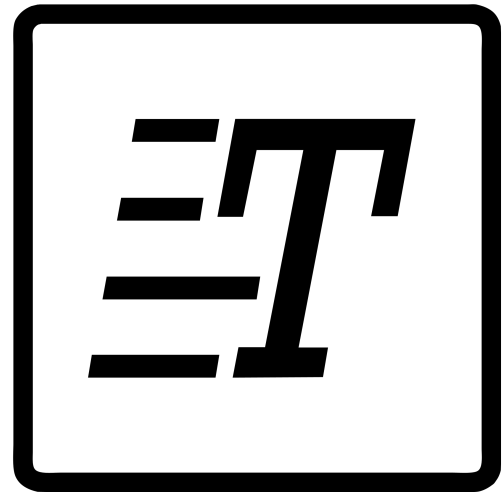
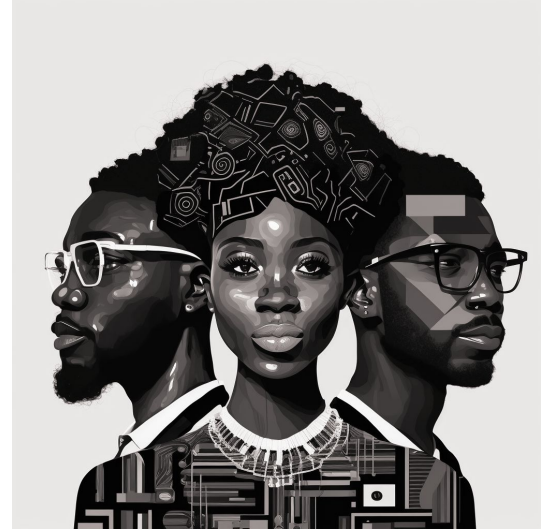
- Tokenization
- Data cleaning (email, URL extraction)
- Normalization (lowercasing, punctuation removal)

Key Regex Patterns:

- `\d+` → Matches digits (123)
- `\w+` → Matches words (hello123)
- `\s+` → Matches spaces

 **Example:** Extract emails from text
python

```
import re
text = "Contact us at info@nlpup.ac.za"
print(re.findall(r"\b[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\.[A-Z|a-z]{2,}\b", text))
```



Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Practical Applications of Regex in NLP

Pattern Matching

- Identify **dates**, **phone numbers**, **currency values**.

Tokenization via Regex

- Split text on **non-alphanumeric characters**.

python

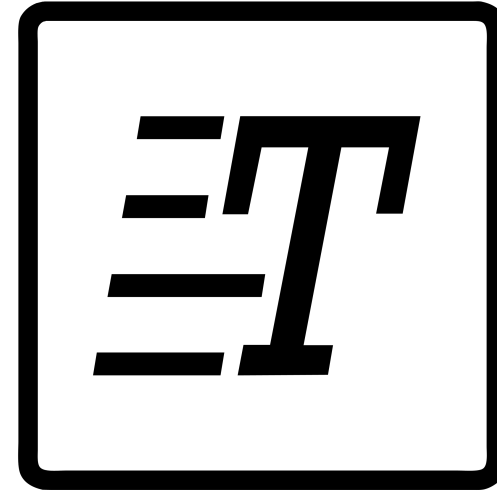
```
import re
text = "Hello, world! Welcome to NLP."
print(re.split(r'\W+', text)) # ['Hello', 'world', 'Welcome', 'to', 'NLP']
```

Cleaning & Preprocessing

- Remove stopwords, HTML tags.

Named Entity Recognition (NER)

- Extract **proper names**, **locations**, **companies** using patterns.



Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

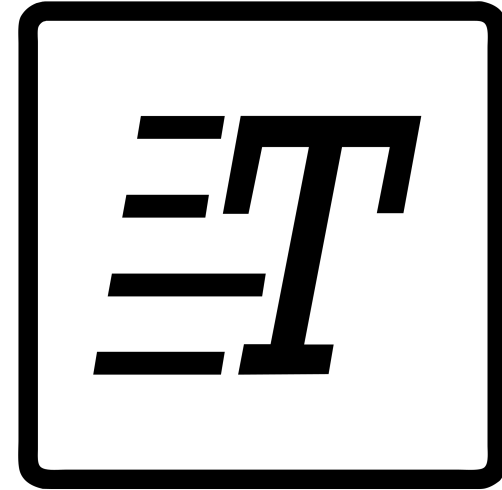
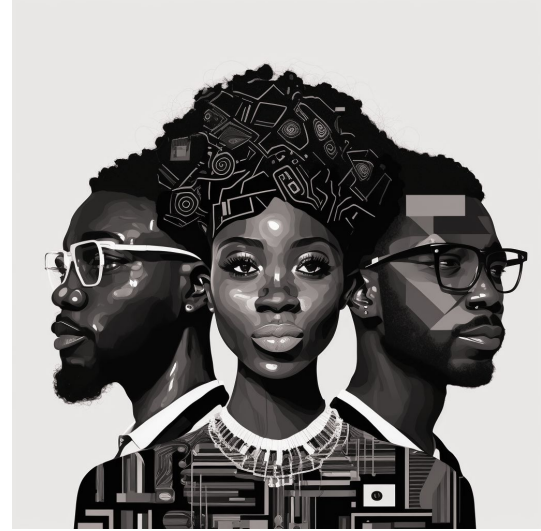
Possibilities

Classification

Sentiment Analysis
Opinion Mining
News categorization
Etc.

Information Retrieval

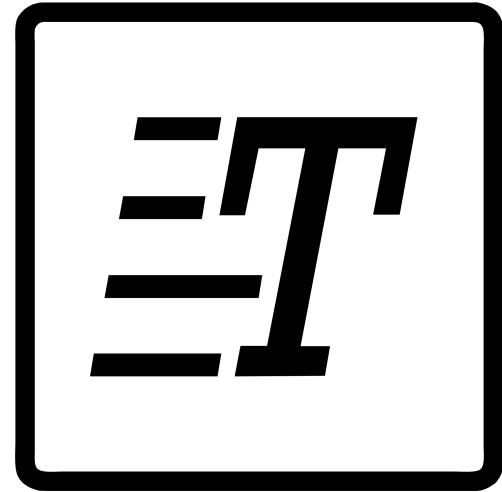
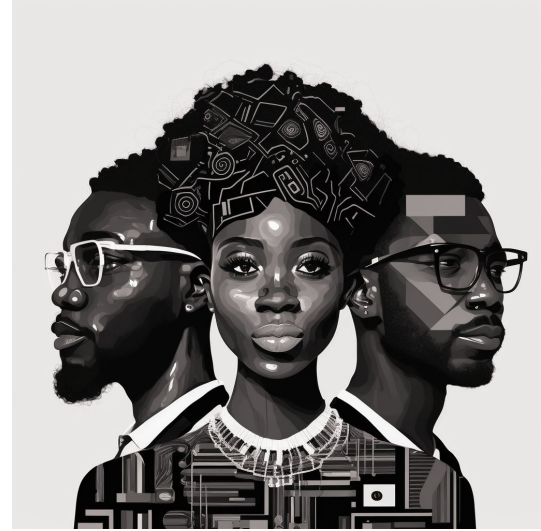
Topic Modelling



Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Challenges



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

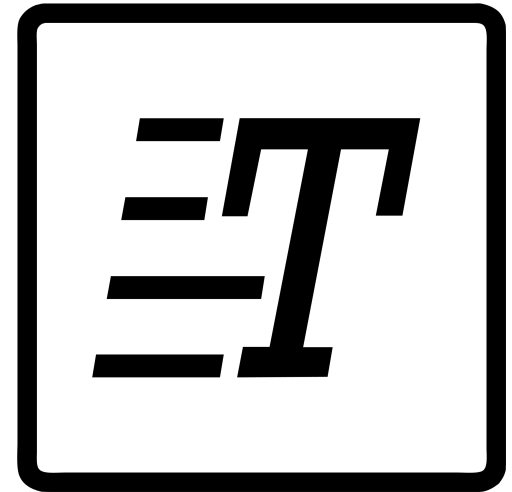
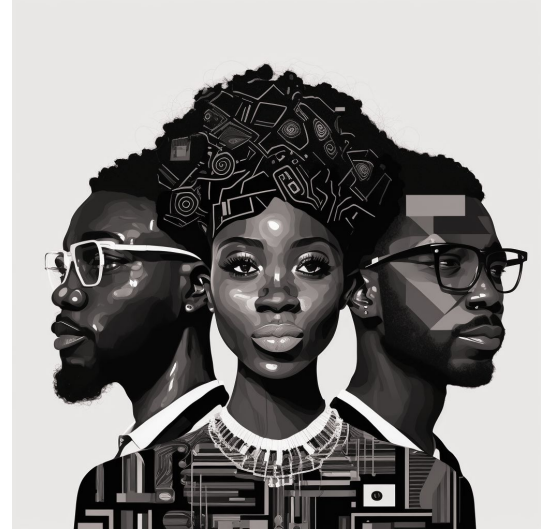
Semantic Meaning

We want terms and phrases that are similar to be treated similarly

Ideas? 💡

- **Synonym generation**
- **Learn a similarity mapping**

We will talk about this in the next lecture.



**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Processes

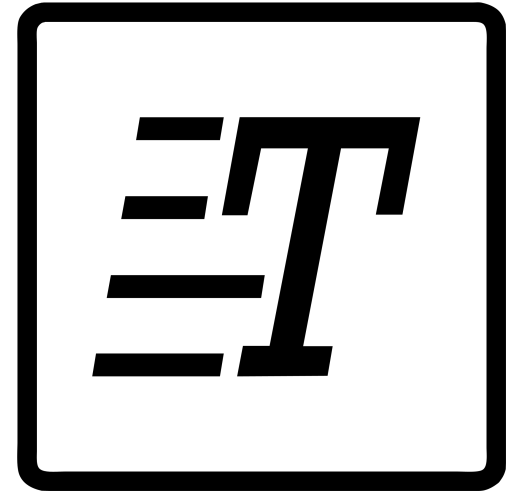
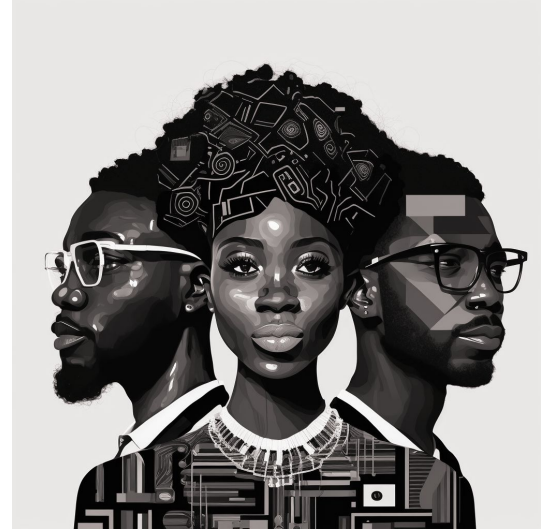
Document -> Symbols [Tokenization]

Symbols -> Vector [Vectorization]

- Frequency
- TF-IDF

Noise?

- Typos
- Misspellings



**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknoloṭši ya Tshedimošo

Language Modelling

- Understanding Language Semantics

Different choices of feature inputs

Estimate:

$$P(w|W) = P(w_{t+1} | w_1, w_2, w_3, w_4)$$

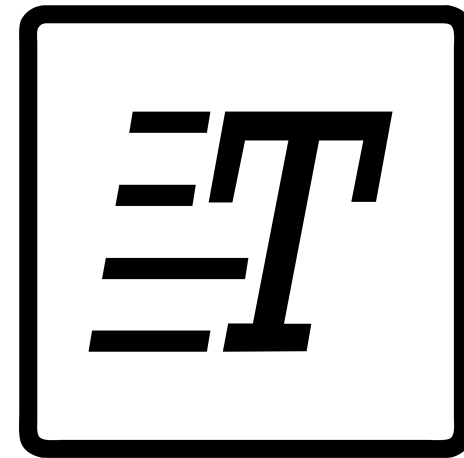
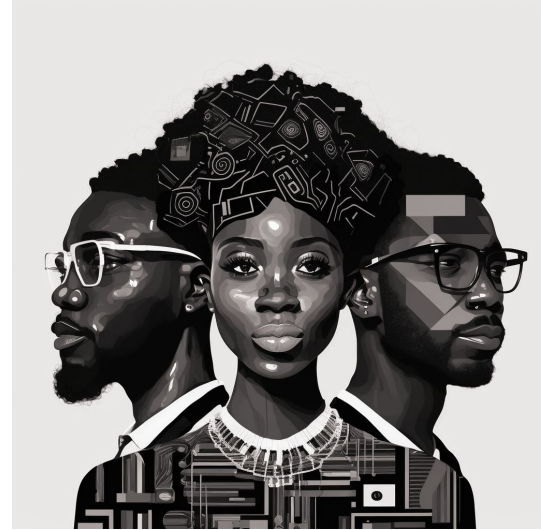
t : next word

W : *word* history/sequence

Can be used for:

- Generating text

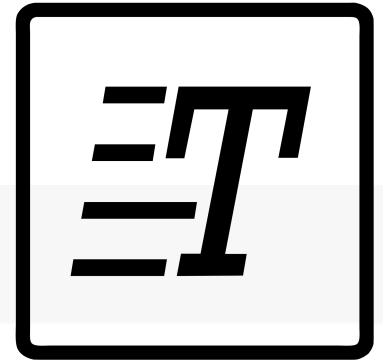
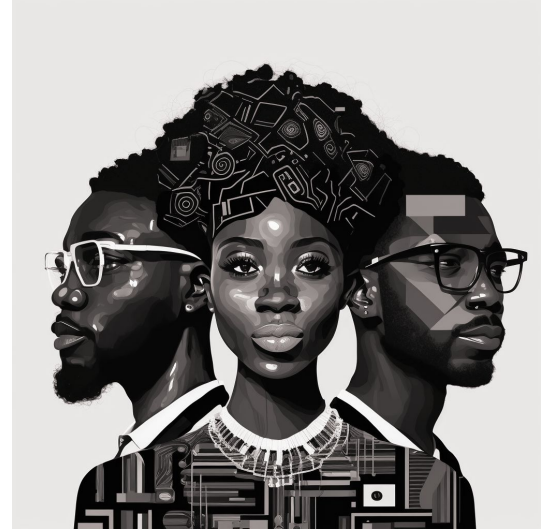
Let's first look at a Statistical Language Model



Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

N-Gram Language Model



Can we calculate the probability of going from a sequence of words to the next word?

- Yes, we can!
- Bonus, we don't need the full sequence [Markov Property]
- We just need a few n-tokens!!!

Demo: simple-ngram-language-model.ipynb

```
[22] 1 print("Probability of text=", prob) # <- Print the probability of the text
      2 print(' '.join([t for t in text if t]))
```

Probability of text= 9.334950669546535e-14

through our industrial strategy and the immense challenge of placing a country is a barometer of the severe economic

Notebook:

<https://colab.research.google.com/drive/1z1eC0kkcAC5J3tCkL8r2tD8fRNMvYFhs?usp=sharing>



Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

N-Gram Language Model

Resources:

Book Chapters

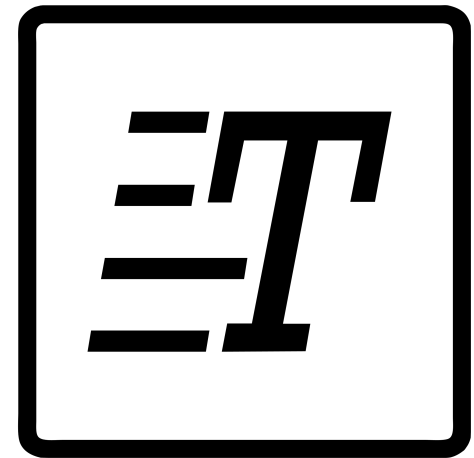
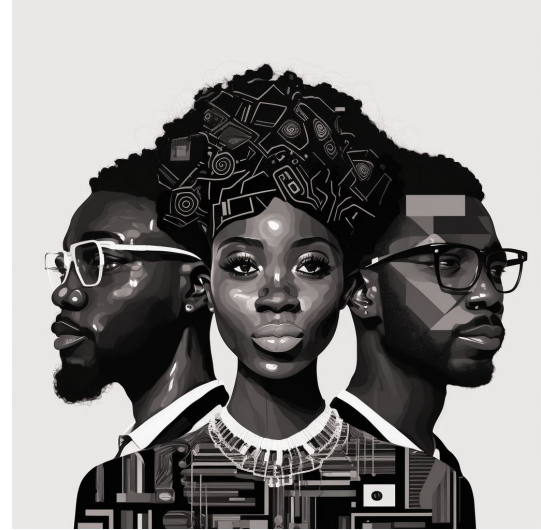
- Speech and Language Processing.
Daniel Jurafsky & James H. Martin [Draft online] -
<https://web.stanford.edu/~jurafsky/slp3/3.pdf>

Slides

- http://www.cs.umd.edu/class/fall2018/cmsc470/slides/slides_10.pdf

Blogs

- <https://nlpforhackers.io/language-models/>
- <https://medium.com/analytics-vidhya/a-comprehensive-guide-to-build-your-own-language-model-in-python-5141b3917d6d>



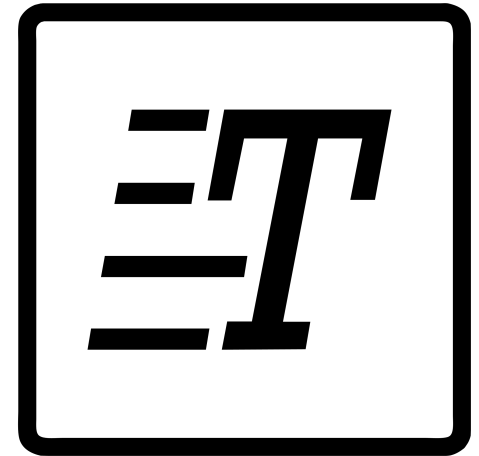
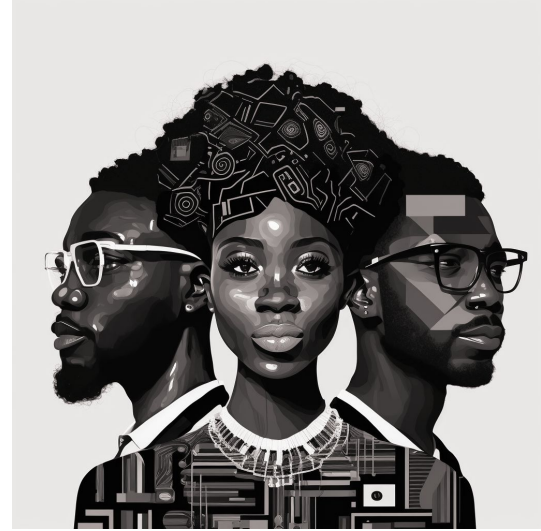
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Recap

- What is NLP?
- Extracting information from text
 - From words to features
 - Bag-Of-Words
 - TFIDF
 - Word Embeddings



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

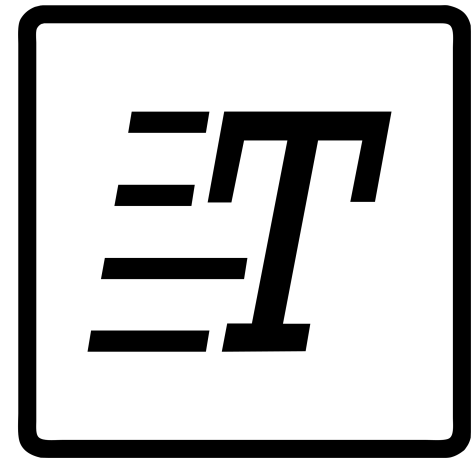
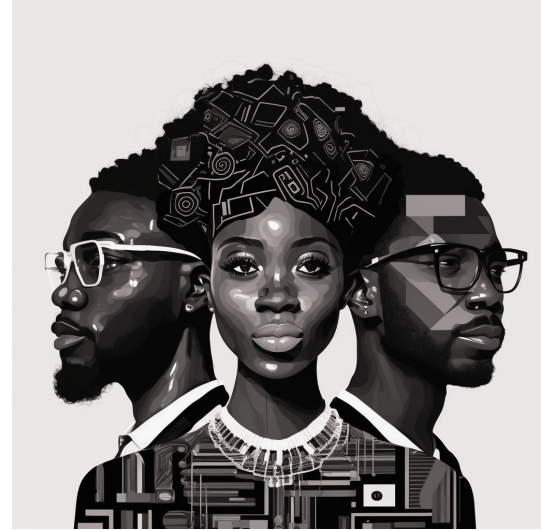
Where to from here?



Assignment: NLP

Jupyter Notebook Based

- Get your feet wet with some NLP tasks
- Will be available from tonight.
 - First NLP assignment subtask is quick
 - Then a few more across the next 2 weeks.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

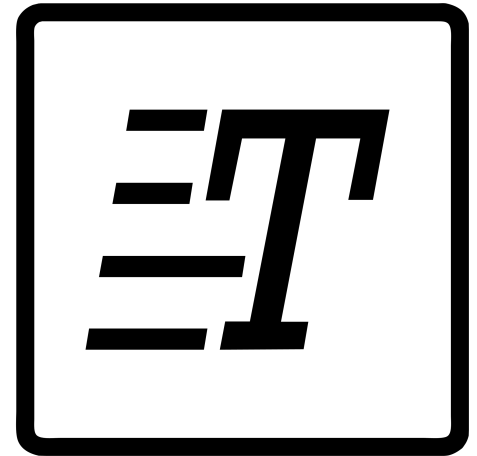
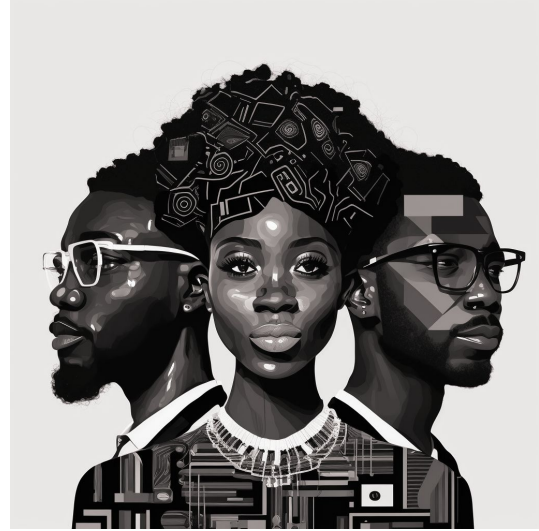
Resources

NLP General

- Fast AI NLP Course <https://github.com/fastai/course-nl>
- Stanford Coursera NLP Slides
<https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html>
- Speech and Language Processing (3rd Ed.)
<https://web.stanford.edu/~jurafsky/slp3/>
- NLP for Hackers <https://nlpforhackers.io/>

Python Libraries

- SKLearn NLP (Working With Text Data) - [URL](#) (Nice tutorial)
- spaCY: Industrial-Strength Natural Language Processing - [URL](#)
- NLTK



Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknološi ya Tshedimošo

Semester Preview

Dr Modupe



Where is the Data? Some Examples

General

- NLP Progress <https://nlpprogress.com/>

Hate Speech

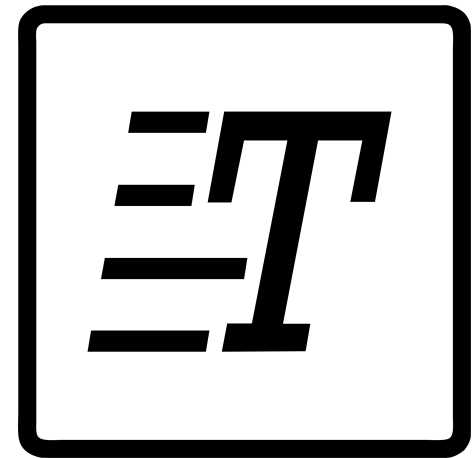
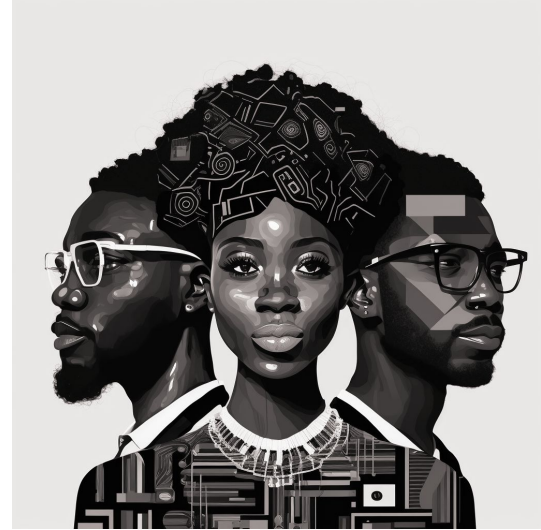
- Hate speech dataset from a white supremacist forum
<https://github.com/aitor-garcia-p/hate-speech-dataset>

Fake News

- Fake News <https://github.com/KaiDMML/FakeNewsNet>
- Real 411 <https://www.real411.org/complaints> [South Africa]

Author Identification

- Style Change
<https://pan.webis.de/clef19/pan19-web/style-change-detection.html>



Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Thank you Questions



Prof. Vukosi Marivate

vukosi.marivate@cs.up.ac.za

<https://dsfsi.github.io>

@vukosi

@DSFSI_Research

Keep in touch

Join our research group newsletter

<https://tinyletter.com/datascience-up/>

Made with ❤️ in Tshwane



**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za