Norms of NLP

Understand and extract knowledge from text data

Applications

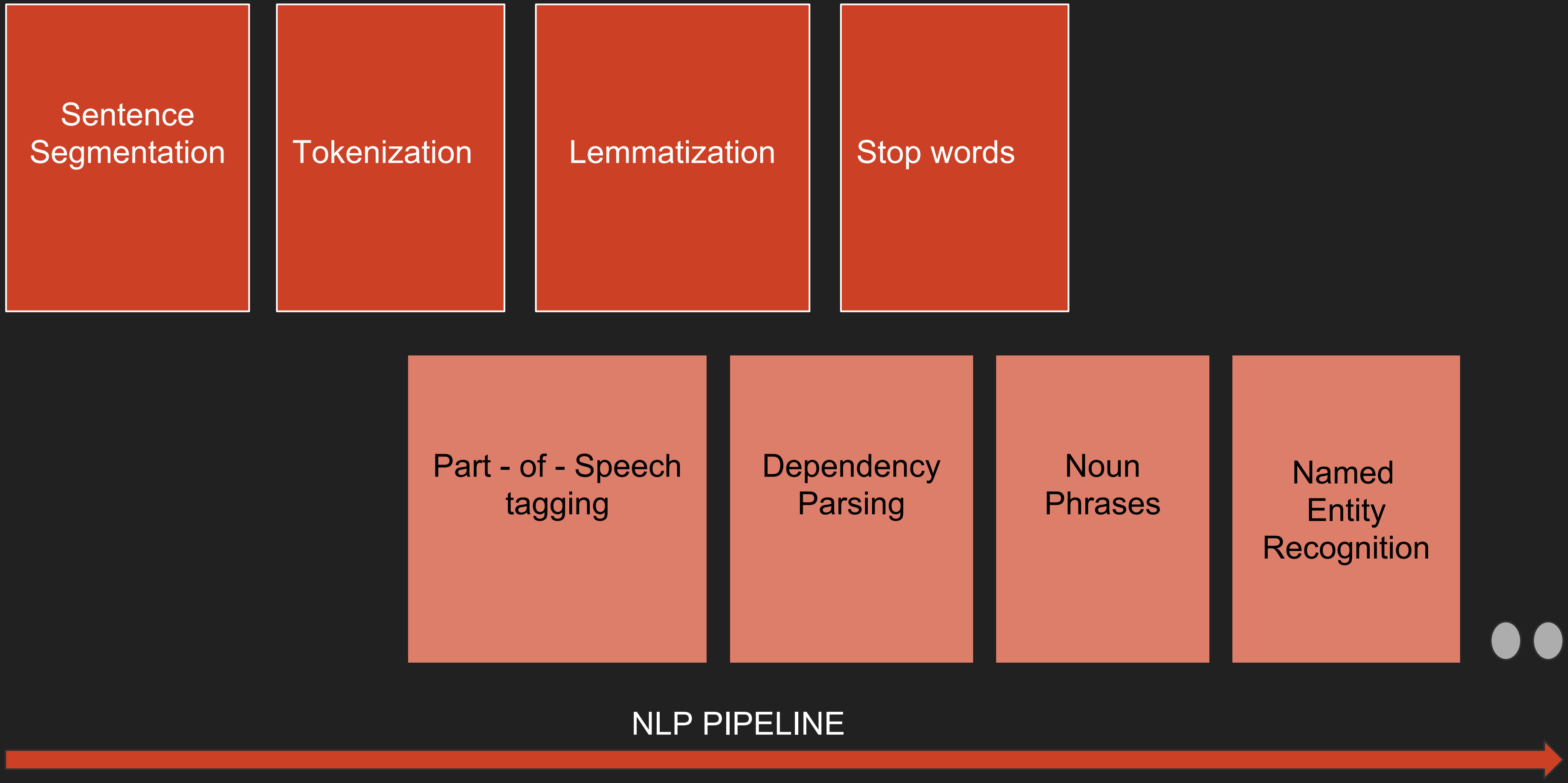
1. Sentiment Analysis

2. Text classification

3. Question Answering

4. Automatic Summarization

5. Machine Translation



Sentence Segmentation

Tokenization

Lemmatization

Stop words

Part - of - Speech tagging

Dependency Parsing

Noun Phrases

Named Entity Recognition

NLP PIPELINE

ELIZA

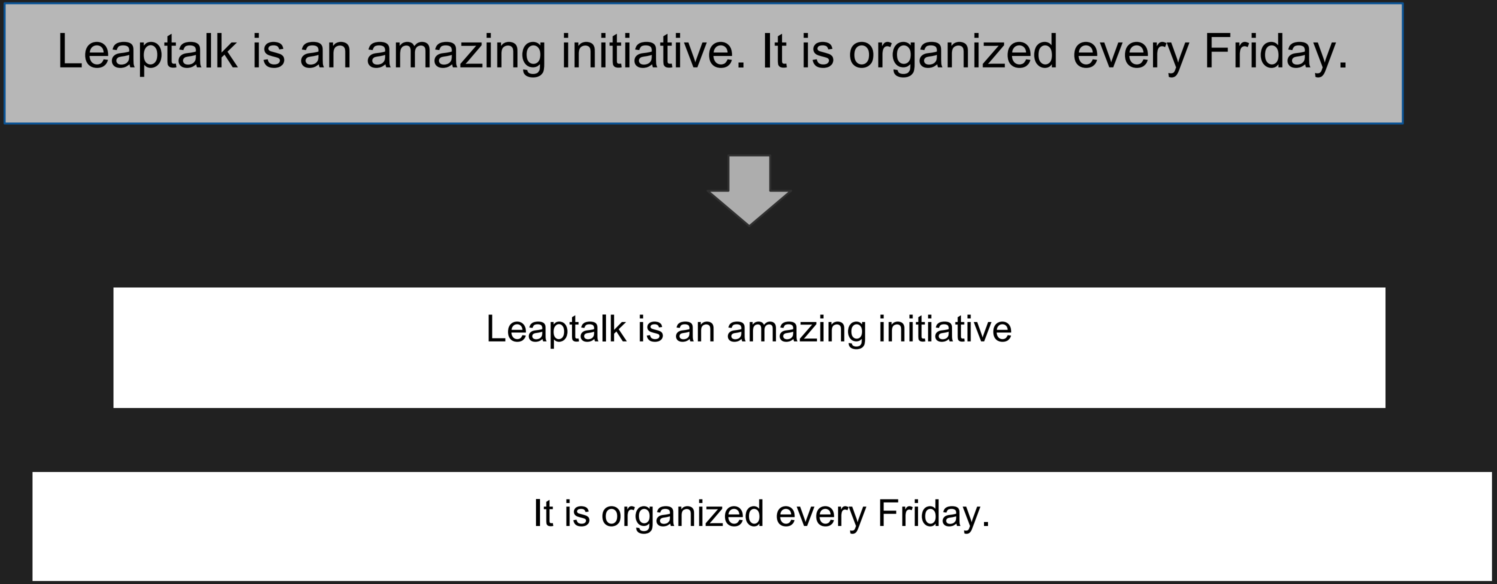
User : You are like my father in some ways.

Eliza : What makes you think I am like your father in some ways?

Text Normalization

1. Sentence Segmentation
2. Tokenization
3. Stopwords
4. Lemmatization -- stemming

Sentence segmentation

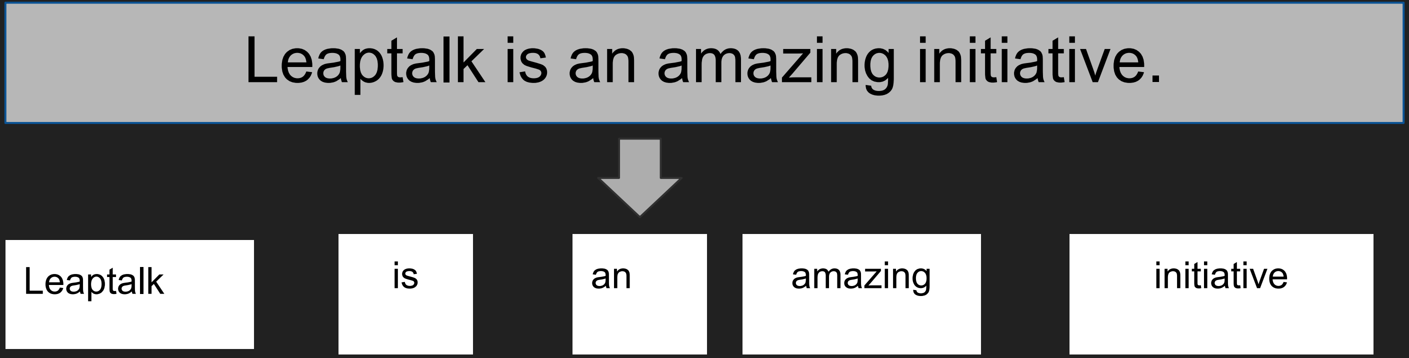


Leaptalk is an amazing initiative. It is organized every Friday.

Leaptalk is an amazing initiative

It is organized every Friday.

Tokenization



Leaptalk is an amazing initiative.

Leaptalk

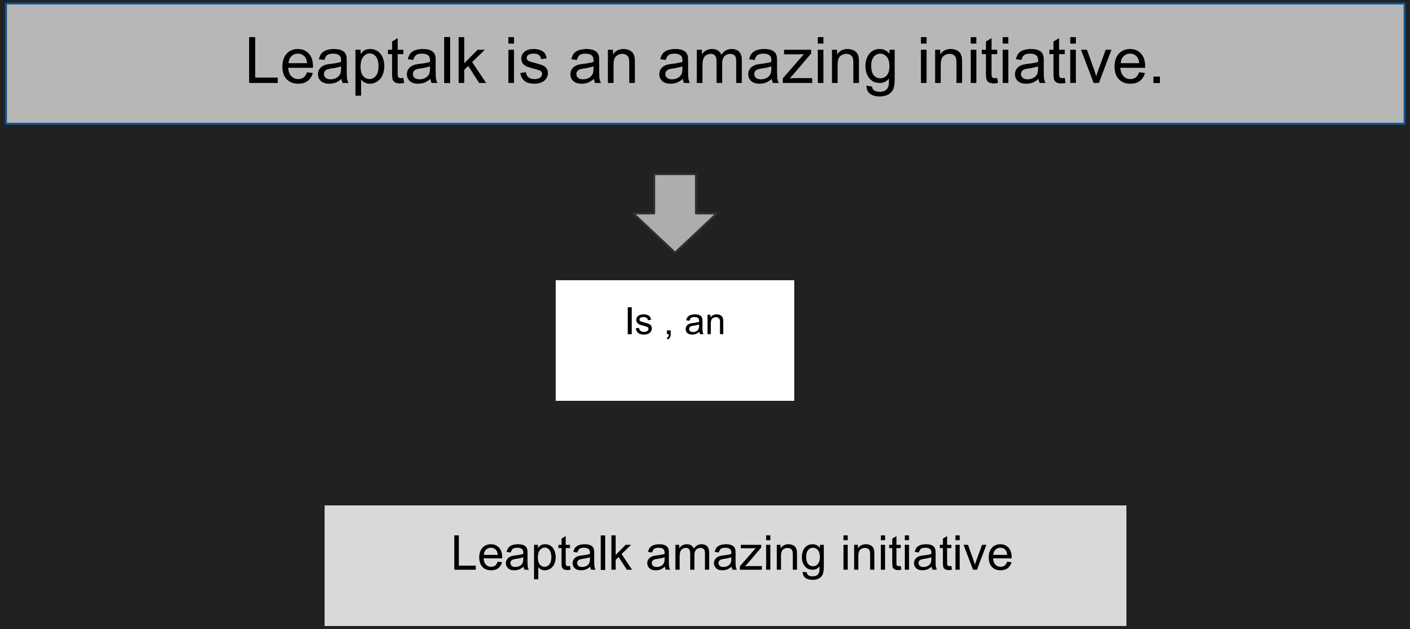
is

an

amazing

initiative

Stopwords

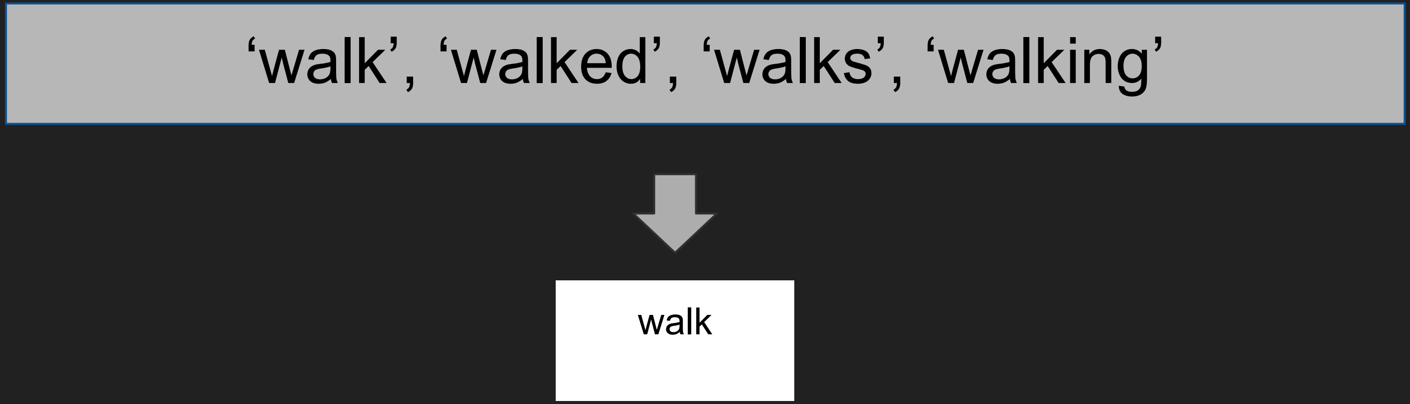


Leaptalk is an amazing initiative.

Is , an

Leaptalk amazing initiative

Lemmatization



‘walk’, ‘walked’, ‘walks’, ‘walking’

walk

Feature Extraction

1. Bag of words (BOW)
2. TF - IDF
3. Cosine similarity
4. Word Mover’s Distance - Looks for semantic meaning

Bag of words (BOW)

D1 : “I am learning NLP”

D2 : “I am learning new things”

Unique : I am learning NLP new things. (6 words)

V1 : (1, 1, 1, 1, 0, 0)   V2 : (1, 1, 1, 0, 1, 1)

	I	am	learning	NLP	new	things
D1	1	1	1	1	0	0
D2	1	1	1	0	1	1

Term Frequency - Inverse Doc. Frequency

Relevance of a term(word) is not proportional to the frequency of terms.

Term	Count
Nepal	1
is	2
a	2
country	1

d1

Term	Count
Nepal	1
is	2
a	2
landlocked	2
country	1

d2

TF = Term occurrence / total count
IDF = log(Total documents / documents with term appearance)

Term Frequency - Inverse Doc. Frequency

Term	Count
Nepal	1
is	2
a	2
country	1

d1

Term	Count
Nepal	1
is	2
a	2
landlocked	2
country	1

d2

For ‘country’

TF (‘country’, d1) = 1 / 6

TF (‘country’, d2) = 1 / 8

IDF (‘country’, D) = log (2 / 2) = 0

TF-IDF = TF * IDF = 0

Term Frequency - Inverse Doc. Frequency

Term	Count
Nepal	1
is	2
a	2
country	1

d1

Term	Count
Nepal	1
is	2
a	2
landlocked	2
country	1

d2

For ‘landlocked’

TF (‘landlocked,’ d1) = 0 / 6 = 0

TF (‘landlocked’, d2) = 2 / 8 = 0.25

IDF (‘landlocked’, D) = log(2/1) = 0.3

TF-IDF (d2) = TF * IDF = 0.25 * 0.3

= 0.075

Cosine similarity

Cosine Similarity (d1, d2) = Dot product(d1, d2) / ||d1|| * ||d2||

Dot product (d1,d2) = d1[0] * d2[0] + d1[1] * d2[1] * … * d1[n] * d2[n]

||d1|| = square root($d1[0]^2$ + $d1[1]^2$ + ... + $d1[n]^2$)

||d2|| = square root($d2[0]^2$ + $d2[1]^2$ + ... + $d2[n]^2$)

Cosine similarity - Illustration

D1 - ‘This is first example for first topic’
D2 - “This is an example for this topic’

Total length : 7   (this is first example for topic an)

D1 : (1, 1, 2, 1, 1, 1, 0)
D2 : (2, 1, 0, 1, 1, 1, 1)

Dot (D1, D2) = 2 + 1 + 0 + 1 + 1 + 1 = 6   and ||d1|| = 3 ||d2|| = 3

Cosine = 6 / (3 * 3) = 0.667

What about semantic?

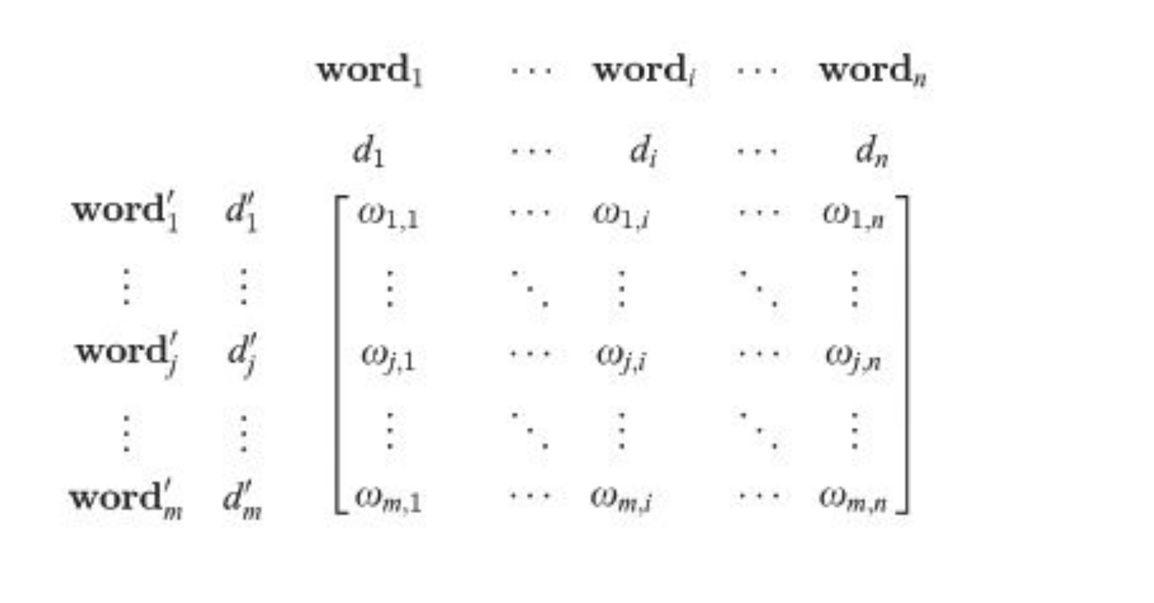
D1 : I speak three languages.

D2 : I know Nepali, English and German.

Cosine similarity would be nearly Zero.

Word Mover’s distance - observe semantic meaning

minimum amount of distance that words of one document need to “travel” to reach the words of another document



$$\begin{array}{ccccccc}
 & & \mathbf{word}_1 & \cdots & \mathbf{word}_i & \cdots & \mathbf{word}_n \\
 & & d_1 & \cdots & d_i & \cdots & d_n \\
\mathbf{word}'_1 & d'_1 & \omega_{1,1} & \cdots & \omega_{1,i} & \cdots & \omega_{1,n} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\mathbf{word}'_j & d'_j & \omega_{j,1} & \cdots & \omega_{j,i} & \cdots & \omega_{j,n} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\mathbf{word}'_m & d'_m & \omega_{m,1} & \cdots & \omega_{m,i} & \cdots & \omega_{m,n}
\end{array}$$

Word Mover’s distance - observe semantic meaning

- Algorithm uses already built word embeddings developed using word2Vec model.
- Word2Vec : algorithm which transforms words to vectors and and words having similar meaning laying close to each other.

King - Man + Woman = Queen