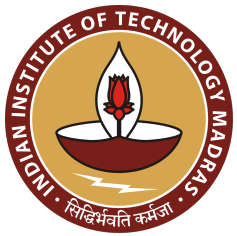
BSCS5002: Introduction to Natural Language Processing

Week 2 Lecture-2: Text Analysis: Stemming and Lemmatization

Parameswari Krishnamurthy



Language Technologies Research Centre
IIIT-Hyderabad

param.krishna@iiit.ac.in

Text Analysis: Linguistic Representation

Stemming

Stemming refers to the process of slicing a word with the intention of removing affixes. Known as crude chopping of affixes.

Stemming

Stemming refers to the process of slicing a word with the intention of removing affixes. Known as crude chopping of affixes.

Stemming is problematic in the linguistic perspective, since it sometimes produces words that are not in the language, or else words that have a different meaning.

Stemming

Stemming refers to the process of slicing a word with the intention of removing affixes. Known as crude chopping of affixes.

Stemming is problematic in the linguistic perspective, since it sometimes produces words that are not in the language, or else words that have a different meaning.

- Language dependent

Example :

Stemming

Stemming refers to the process of slicing a word with the intention of removing affixes. Known as crude chopping of affixes.

Stemming is problematic in the linguistic perspective, since it sometimes produces words that are not in the language, or else words that have a different meaning.

- Language dependent

Example :

- arguing $>$ argu, flies $>$ fli
- playing $>$ play, caring $>$ car
- news $>$ new

Examples of Stemming

- **Original Word → Stemmed Form**
  - “Caring” → “car”
  - “Studied” → “studi”
  - “Running” → “runn”
  - “Happiness” → “happi”

Advantages and Disadvantages of Stemming

- Advantages of Stemming
  - Fast and simple to implement.
  - Reduces dimensionality of text data, making it easier to analyze.

Advantages and Disadvantages of Stemming

- Advantages of Stemming
  - Fast and simple to implement.
  - Reduces dimensionality of text data, making it easier to analyze.
- Disadvantages of Stemming
  - Sometimes too aggressive, leading to non-words.
  - Example: “studies” $\rightarrow$ “studi”

Advantages and Disadvantages of Stemming

- Advantages of Stemming
  - Fast and simple to implement.
  - Reduces dimensionality of text data, making it easier to analyze.
- Disadvantages of Stemming
  - Sometimes too aggressive, leading to non-words.
  - Example: “studies” $\rightarrow$ “studi”
  - May result in words that lose their meaning.
  - Example: “caring” $\rightarrow$ “car”

Lemmatization

Lemmatization has the objective of reducing a word to its base form, also called Lemma, therefore grouping together different forms of the same word.

Lemmatization

Lemmatization has the objective of reducing a word to its base form, also called Lemma, therefore grouping together different forms of the same word.

- Have to find correct dictionary headword form

Lemmatization

Lemmatization has the objective of reducing a word to its base form, also called Lemma, therefore grouping together different forms of the same word.

- Have to find correct dictionary headword form

Example :

- am, are, is $>$ be

Lemmatization

Lemmatization has the objective of reducing a word to its base form, also called Lemma, therefore grouping together different forms of the same word.

- Have to find correct dictionary headword form

Example :

- am, are, is $>$ be
- car, cars, car's, cars' $>$ car

Lemmatization

Lemmatization has the objective of reducing a word to its base form, also called Lemma, therefore grouping together different forms of the same word.

- Have to find correct dictionary headword form

Example :

- am, are, is $>$ be
- car, cars, car's, cars' $>$ car
- the boy's cars are different colors $>$ the boy car be different color

Lemmatization

Lemmatization has the objective of reducing a word to its base form, also called Lemma, therefore grouping together different forms of the same word.

- Have to find correct dictionary headword form

Example :

- am, are, is $>$ be
- car, cars, car's, cars' $>$ car
- the boy's cars are different colors $>$ the boy car be different color

Lemmatization and stemming are mutually exclusive, and the former is much more resource-intensive than the latter.

How It Works?

- Lemmatization requires a detailed understanding of the word's part of speech (POS) and context to transform it into the correct lemma.

How It Works?

- Lemmatization requires a detailed understanding of the word's part of speech (POS) and context to transform it into the correct lemma.
- Examples of Lemmatization
- **Original Word $\rightarrow$ Lemma**
  - "Caring" $\rightarrow$ "care"

How It Works?

- Lemmatization requires a detailed understanding of the word's part of speech (POS) and context to transform it into the correct lemma.
- Examples of Lemmatization
- **Original Word $\rightarrow$ Lemma**
  - "Caring" $\rightarrow$ "care"
  - "Studies" $\rightarrow$ "study"

How It Works?

- Lemmatization requires a detailed understanding of the word's part of speech (POS) and context to transform it into the correct lemma.
- Examples of Lemmatization
- **Original Word → Lemma**
  - "Caring" → "care"
  - "Studies" → "study"
  - "Running" → "run"

How It Works?

- Lemmatization requires a detailed understanding of the word's part of speech (POS) and context to transform it into the correct lemma.
- Examples of Lemmatization
- **Original Word → Lemma**
  - "Caring" → "care"
  - "Studies" → "study"
  - "Running" → "run"
  - "Better" → "good"

How It Works?

- Lemmatization requires a detailed understanding of the word's part of speech (POS) and context to transform it into the correct lemma.
- Examples of Lemmatization
- **Original Word → Lemma**
  - "Caring" → "care"
  - "Studies" → "study"
  - "Running" → "run"
  - "Better" → "good"

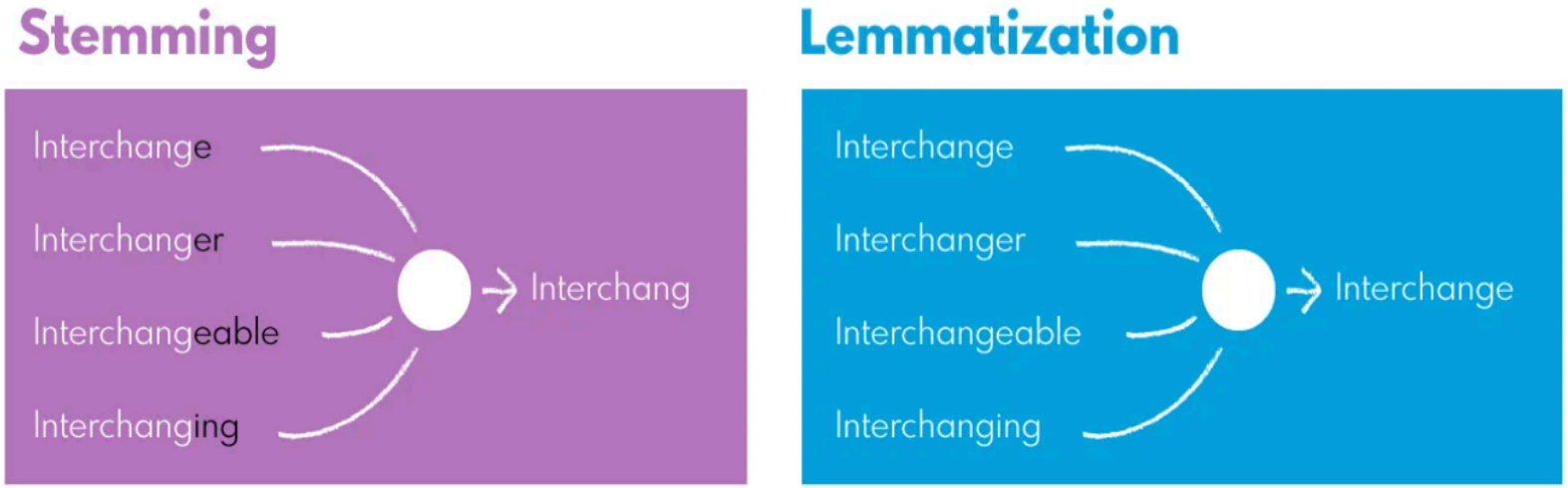
Advantages and Disadvantages of Lemmatization

- Advantages of Lemmatization:
  - More accurate than stemming because it produces real words.
  - Maintains the meaning and grammatical correctness of words.

Advantages and Disadvantages of Lemmatization

- Advantages of Lemmatization:
  - More accurate than stemming because it produces real words.
  - Maintains the meaning and grammatical correctness of words.
- Disadvantages of Lemmatization:
  - Slower and more complex to implement.
  - Requires additional resources like a dictionary or POS tagger.

Stemming Vs. Lemmatization



Stemming vs Lemmatization

Comparison of Stemming and Lemmatization

Feature	**Stemming**	**Lemmatization**
Approach	Rule-based, chops off suffixes	Dictionary-based, considers POS
Result	Stem (may not be a real word)	Lemma (always a real word)
Speed	Faster, less computationally intensive	Slower, more computationally intensive
Accuracy	Less accurate, may distort meaning	More accurate, preserves meaning
Use Case	Simple text processing	Advanced text analysis and NLP tasks
"Running"	Stem: "runn"	Lemma: "run"
"Studies"	Stem: "studi"	Lemma: "study"

Practical Applications of Stemming

- **Search Engines:**
  - Reduces variations of words to their base form to improve search results.
  - For example, searching for "run" might also return "running" and "ran".
- **Text Mining:**
  - Simplifies words in a large dataset, making it easier to analyze patterns.

Practical Applications of Lemmatization

- **Machine Translation:**
  - Ensures that words are translated accurately by maintaining their base form.
- **Sentiment Analysis:**
  - Improves the accuracy of text sentiment analysis by understanding the correct form of words.
- **Speech Recognition:**
  - Helps in identifying the correct form of spoken words to improve transcription accuracy.