National Research
Tomsk
State
University

# Natural Language Processing. Basics. Part 1

## (A Very Short Introduction)

Sergey V. Axyonov,

PhD, Associate Professor,

Department of Fundamental Computer Science, Institute of Applied Math & Computer Science,

Tomsk State University

Tomsk-2024

# Natural Language Processing Tasks

| Task | Description |
|---|---|
| Tokenization | Dividing a text corpus into indivisible units |
| Word disambiguation | Determining the correct meaning of a word |
| Named entity recognition | Extracting entities (names, companies, cities, etc.) from a text corpus |
| Morphological labelling | Identifying parts of speech in a sentence and annotating them |
| Sentence classifier | Assigning texts to certain classes |
| Language generation | Generating new texts using examples |
| Question and answer solutions | Chatbots, information retrieval and knowledge representation |
| Machine translate | Converting texts from one language to another |

# Text Data Encoding: Definition

The process of converting natural language texts into numeric values that computers can understand.

Input: Blood is composed of blood cells suspended in blood plasma.

Encoding: 'blood' → 0, 'is' → 1, 'composed' → 2, 'of' → 3, 'cells' → 4, 'suspended' → 5, 'in' → 6, 'plasma' → 7

Output: 0, 1, 2, 0, 3, 4, 5, 6, 0, 7

# Tokenization: Example

“From my earliest youth I realized that my nature was a mass of contradictions.”
(“And Then There Were None” by Agatha Christie)

1. Letters: 'F', 'r', 'o', 'm', ' ', 'm', 'y', ' ', 'e', 'a', 'r', 'l', 'i', 'e', 's', 't’, …

2. Subwords: 'Fr', 'o', 'm', ' ', 'my', ' ', 'ea', 'r', 'l', 'ie', 'st’, …

3. Words: 'From', 'my', 'earliest', 'youth', 'I', 'realized', 'that', 'my', 'nature’, …

4. N-grams (Some neighboring words): 'From my', 'my earliest', 'earliest youth', 'youth I', 'I realized', 'realized that', 'that my', 'my nature', …

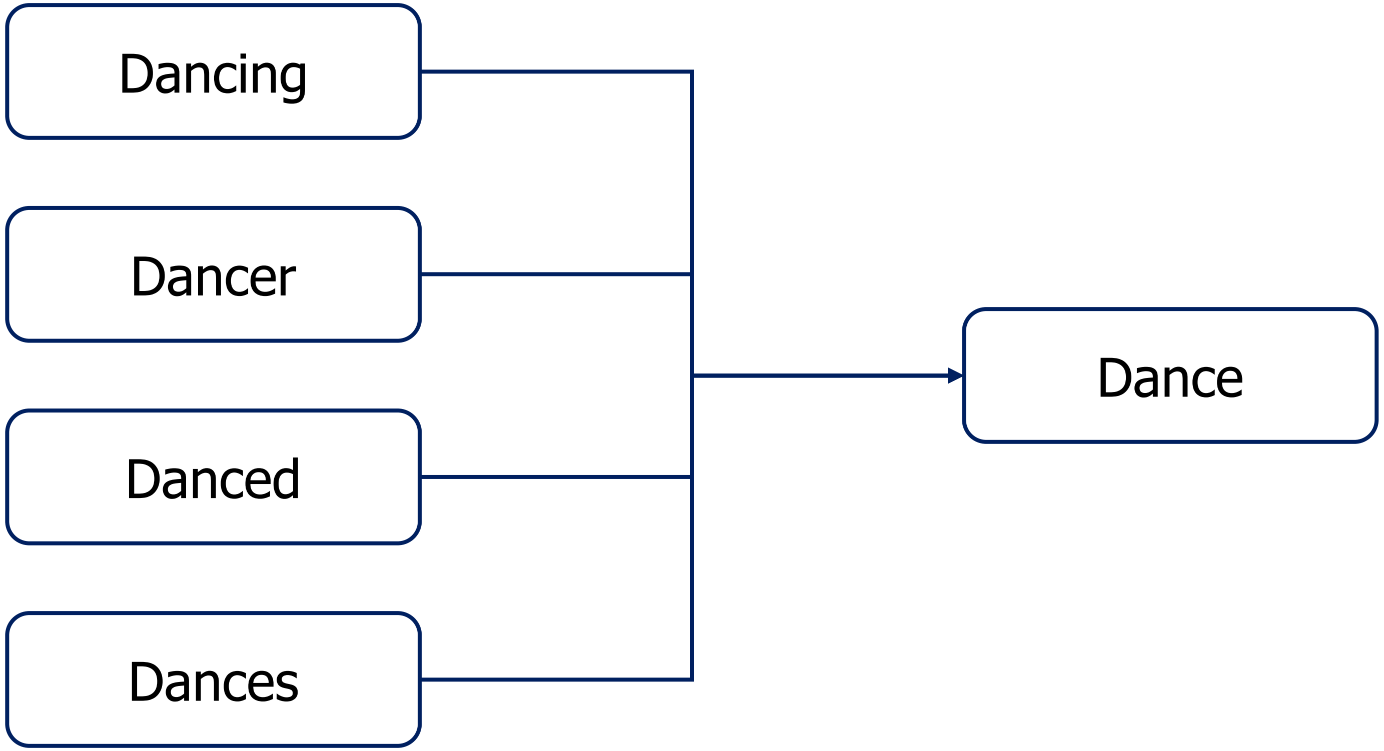
# Some Issues (but not all)

1. Multiple meanings: "back" (Noun, Verb, Adjective), "date" (Noun, Verb), "fair" (Noun, Adjective), "present" (Noun, Verb, Adjective), …

2. Synonyms: "conversation" ↔ "talk", "amazing" ↔ "incredible", "begin" ↔ "start", "help" ↔ "aid“

3. Word formation: "reason" → "reasonable", "happy" → "unhappy", "nature" → "natural" → "naturally" → "unnaturally"

# Reducing Word Variants to One Base Form



# Stemming: Definition

✓ Stemming algorithms eliminate word suffixes by running input word tokens against a pre-defined list of common suffixes.

✓ The stemmer then removes any found suffix character strings from the word, should the latter not defy any rules or conditions attached to that suffix.

✓ Some stemmers run the resulting stemmed bits through an additional set of rules to correct for malformed roots.

Source: https://www.ibm.com/topics/stemming-lemmatization

# Stemming: Examples

Examples:

1. “Life appears to me too short to be spent in nursing animosity or registering wrongs.” (“Jane Eyre” by Charlotte Brontë)

2. “I am the happiest creature in the world. Perhaps other people have said so before, but not one with such justice.” (“Pride and Prejudice” by Jane Austen)

Stemmed words:

1. 'life', 'appear', 'to', 'me', 'too', 'short', 'to', 'be', 'spent', 'in', 'nurs', 'animos', 'or', 'regist', 'wrong'

2. 'i', 'am', 'the', 'happ', 'creatur', 'in', 'the', 'world', 'perhap', 'other', 'peopl', 'have', 'said', 'so', 'befor', 'but', 'not', 'one', 'with', 'such', 'justic'

# Lemmatization: Definition

✓ Lemmatization is the larger enterprise of reducing morphological variants to one dictionary base form.

✓ The practical distinction between stemming and lemmatization is that, where stemming merely removes common suffixes from the end of word tokens, lemmatization ensures the output word is an existing normalized form of the word (for example, lemma) that can be found in the dictionary.

✓ Because lemmatization aims to output dictionary base forms, it requires more robust morphological analysis than stemming.

Source: https://www.ibm.com/topics/stemming-lemmatization

# Lemmatization

Examples:

1. “Life appears to me too short to be spent in nursing animosity or registering wrongs.” (“Jane Eyre” by Charlotte Brontë)

2. “I am the happiest creature in the world. Perhaps other people have said so before, but not one with such justice.” (“Pride and Prejudice” by Jane Austen)

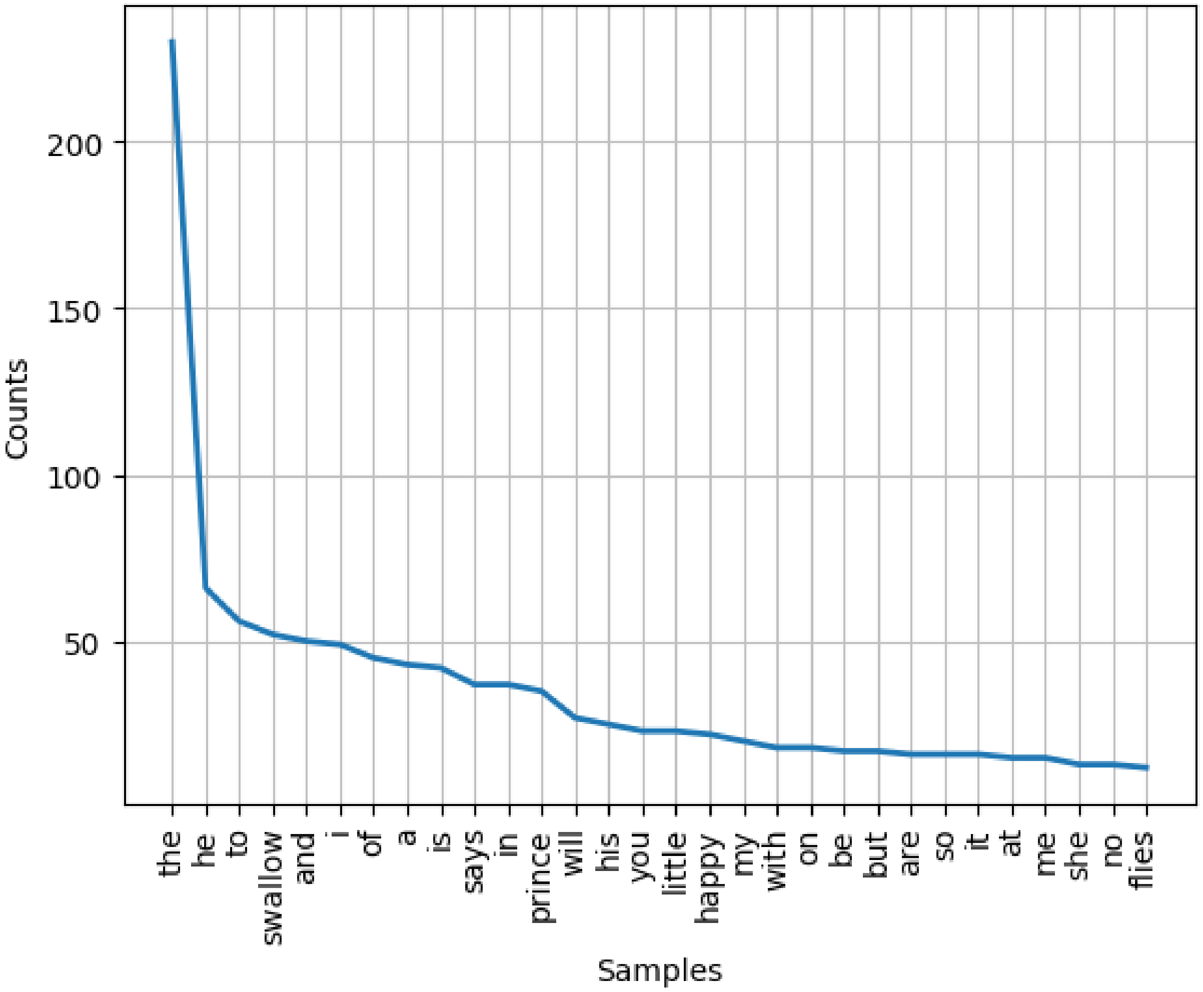
Lemmatized words:

1. 'life', 'appear', 'to', 'me', 'too', 'short', 'to', 'be', 'spend', 'in', 'nursing', 'animosity', 'or', 'registering', 'wrong'

2. 'i', 'am', 'the', 'happy', 'creature', 'in', 'the', 'world', 'perhaps', 'other', 'people', 'have', 'say', 'so', 'before', 'but', 'not', 'one', 'with', 'such', 'justice'

# Frequency Analysis of Text: Example

"The Happy Prince" by Oscar Wilde

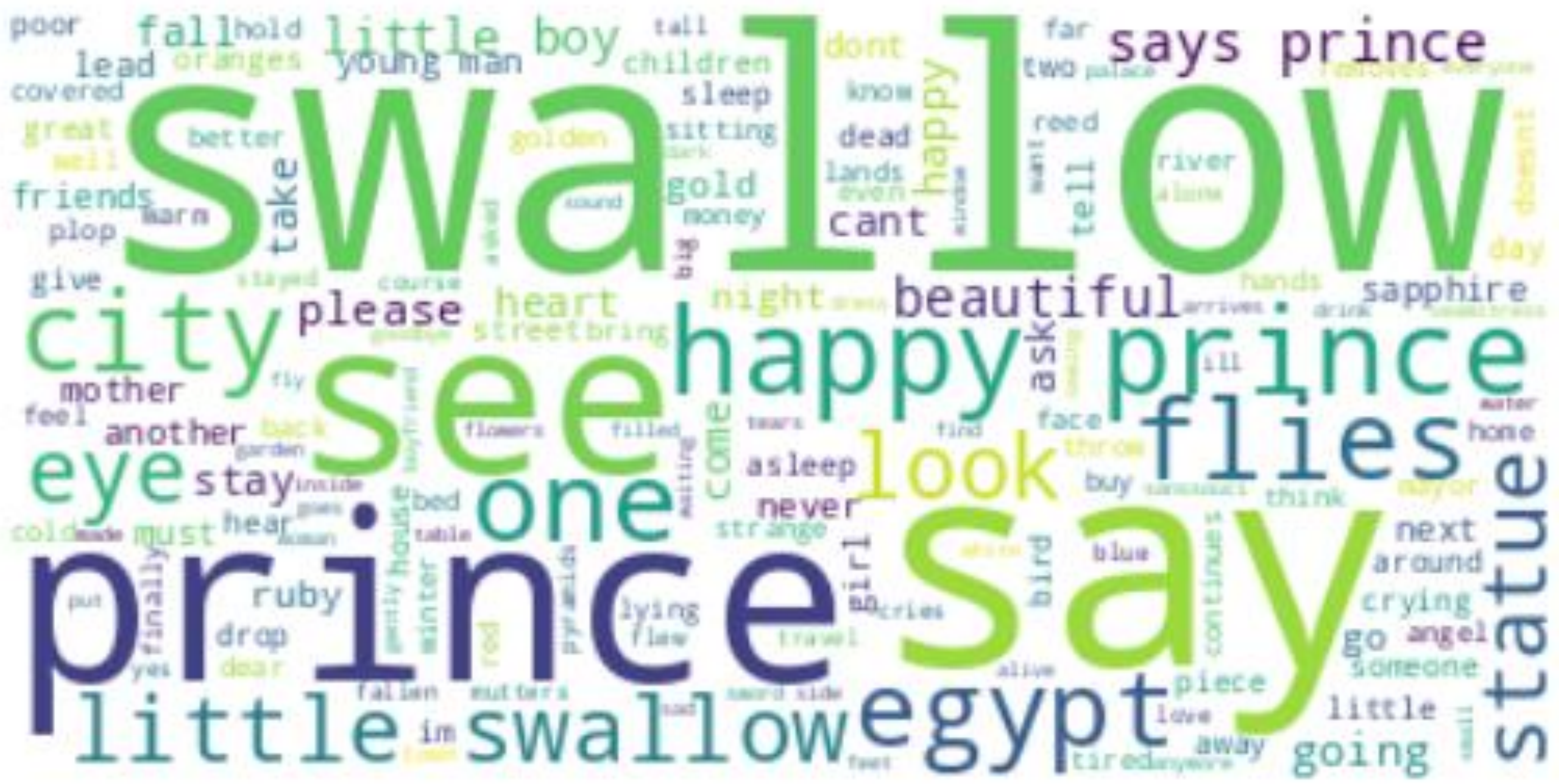


Top 10 words after stop-words removing:

```
('swallow', 52),
('says', 37),
('prince', 35),
('little', 23),
('happy', 22),
('flies', 12),
('egypt', 12),
('one', 12),
('city', 11),
('statue', 10)
```

# Word Cloud: Example

“The Happy Prince” by Oscar Wilde



# One-Hot Encoding

Input: 'Blood is composed of blood cells suspended in blood plasma'.

Stop-words: 'is', 'of', 'in'.

Indexing: 'blood' → 0, 'compose' → 1, 'cell' → 2, 'suspend' → 3, 'plasma' → 4

| Token | | | | | |
|---|---|---|---|---|---|
| blood | 1 | 0 | 0 | 0 | 0 |
| compose | 0 | 1 | 0 | 0 | 0 |
| cell | 0 | 0 | 1 | 0 | 0 |
| suspend | 0 | 0 | 0 | 1 | 0 |
| plasma | 0 | 0 | 0 | 0 | 1 |

# Bag-of-Words: Example

Vocabulary: 'blood', 'cell', 'plasma', 'presence', 'computer', 'quantum', 'mechanical', 'compose', 'take'

1. Blood is composed of blood cells suspended in blood plasma.
2. Red blood cells and white blood cells make about forty percent of the blood.
3. The presence of the red blood cells gives the blood a deep-red shade.
4. A quantum computer is a computer that takes advantage of quantum mechanical phenomena.
5. A quantum computer leverages quantum superposition and entanglement.

| blood | cell | plasma | presence | computer | quantum | mechanical | compose | take |
|---|---|---|---|---|---|---|---|---|
| 3 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 2 | 2 | 1 | 0 | 1 |
| 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 |