# титулка

# Зміст

Анотація

Introduction to natural language processing

Linguistic background

Mathematical models

Python implementation

Appendix

# анотація

# 1. Вступ в обробку натуральних мов

Computers se text that they don’t really understand

They have to use some prior knowledge.

They reason probabilistically.

They use context.

# 2. Лінгвістична база

The candidate is preparing for his run for the presidency.

The swimmer is getting ready to run in the final race.

**run** — verb or noun?

## Parts of speech

Penn treebank tagset

need some observations of ambiguity e.g.: 11% of all types but 40% of all tokens in the Brown corpus are ambiguous.

### Brown corpus

History

Regular expression for tag separation

## Stemming

## Tokenization

Example grammars?

# 3. Математичні моделі

## Probabilities

Conditional probability

chain rule

random variables

bayes’ theorem

probabilistic language model

* Speech recognition
  + P(“recognize speech”) > P(“wreck a nice beach”)
* Text generation
  + P(“three houses”) > P(“three house”)
* Spelling correction
  + P(“my cat eats fish”) > P(“my xat eats fish”)
* Machine translation
  + P(“the blue house”) > P(“the house blue”)
* Other uses
  + OCR
  + Summarization
  + Document classification

probability of a sentence, chain rule

## N-grams

Markov assumtption

random text example

optimal N?

Maximum likelihood estimation

n-grams and weighted regular languages

## Markov models

examples

## Hidden Markov models

Examples

viterbi algorithm

supervised learning

# 4. Реалізація мовою Python

## Procedural

Few functions

brevity

keep it simple!

### Documentation

Pyhon має особливу та універсальну підтримку документування функцій та класів. Якщо перший вираз у тілі об’єкта або функції складається лише з одного рядкового літерала, то він вважається рядком документації (docstring). Звернення до рядка документації виконується у зручний спосіб за допомогою функції help.

Наприклад, документація до функції math.sin виглядає наступним чином:

Help on built-in function sin in module math:

sin(...)

sin(x)

Return the sine of x (measured in radians).

## Dictionaries

Dictonaries are hash-tables

allow to use complex-type indices like tuples of strings (n-gram representation)

## NLTK

Natural language tool kit

## N-grams

Useful functions

## Native tagger?

Worse :)

## Pickle

Pickle allows object serialization

# 5. Додаток

# ? u sure ?