УДК 004.622

В.В. Яцишин канд. техн. наук, А.О. Давидов, Д.О. Подолян
Тернопільський національний технічний університет імені Івана Пулюя, Україна

КЛАСИФІКАЦІЯ ТА ПРЕПРОЦЕСИНГ ТЕКСТОВИХ ДАНИХ

V.V. Yatsyshyn PhD, Assoc. Prof., A.O. Davydov, D.O. Podolyan

TEXTS CLASSIFICATION AND PREPROCESSING

На сучасному етапі розвитку штучного інтелекту запропоновано багато методів для вирішення задач класифікації текстів за допомогою автоматичних процедур. Основне призначення таких методів – аналіз, класифікація та виявлення прихованих закономірностей у великих обсягах різнорідних складно структурованих даних. Існуючі методи доцільно поділити на два принципово різних класи: методи машинного навчання і методи, засновані на знаннях (так званий “інженерний підхід”).

Однією із задач, які передують безпосередній класифікації тексту, є препроцесинг, який включає так звану токенізацію тексту. Задача токенізації полягає у розбитті тексту на слова, які називають токенами, з можливим видаленням спеціальних символів, зокрема символів пунктуації. Приклад. Нехай задано вхідну послідовність англійською мовою: «Friends, Romans, Countrymen, lend me your ears». Після проведення токенізації одержуємо вихідну послідовність, як показано на рис. 1.

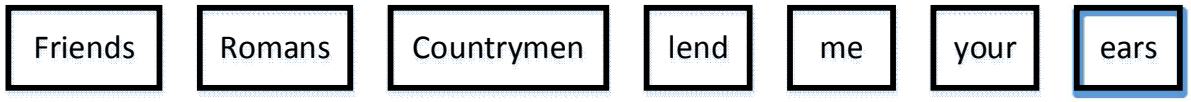

Рисунок 1. Токенізований текст

Токени досить часто називають як терміни або слова, але іноді важливо знати відмінності між типами токенів. Токен – це екземпляр послідовності символів у частині документу, що згруповані разом для зручного використання при семантичному опрацюванні тексту. Тип – це клас всіх токенів, що містить послідовність однакових символів. Терм (може бути нормалізованим) – це тип, що включений у словник системи пошуку інформації (Information retrieval system).

Множина індексів термів може бути цілком відмінною від токенів, для прикладу, вони можуть бути семантичними ідентифікаторами в таксономії, але на практиці в сучасних пошукових системах інформації, вони сильно залежні від токенів у документі. Однак, якщо говорити точніше, токени, що з’являються у документі зазвичай походять від термів шляхом застосування різних підходів до нормалізації.

Найбільш важливе питання процесу токенізації полягає в коректності використання токенів. У попередньому прикладі, процес токенізації є тривіальним, оскільки з речення видалено лише пробіли і знаки пунктуації. Однак в англійській мові та в інших мовах є багато складних випадків. Для прикладу, що робити з апострофом в англійській мові, коли він означає присвійний займенник або скорочення.