УДК 004.91+811.1
А.М. Луцків[1], канд.техн.наук, доц., Н.М. Попович[2], канд.філол.наук, доц., Х.Б. Юркевич[1]
[1]Тернопільський національний технічний університет ім. Івана Пулюя, Україна,
[2]Ужгородський національний університет, Україна

БІБЛІОТЕКИ ОБРОБКИ ПРИРОДНИХ МОВ У ПРЕДМЕТНІЙ ОБЛАСТІ ВЕЛИКИХ ДАНИХ

A.M. Lutskiv (Ph.D.; Assoc. Prof.), N.M. Popovych (Ph.D.; Assoc. Prof.), Kh.B. Yurkevych

NATURAL LANGUAGE PROCESSING LIBRARIES IN A BIG DATA SUBJECT AREA

Опрацювання природної мови людини — NLP (*англ.* Natural Language Processing) є однією із найпоширеніших задач у сучасних інформаційних системах. Сферами застосування NLP[1] є системи автоматичних й автоматизованих перекладів, інформаційно-пошукові системи, засоби корпусної лінгвістики, системи розпізнавання усної мови, системи синтезу мови та багато інших. До базових задач NLP належать:

- визначення належності до частин мови (part-of-speech tagging);
- визначення власних імен сутностей (NER — named-entity recognition);
- машинний переклад;
- синтаксичний поділ на слова та речення;
- категоризація тексту;
- автоматичне визначення мови;
- побудова *n*-грам;
- лематизація та стемінг.

Серед задач опрацювання природної мови людини можна умовно виділити підкатегорію — NLU (*англ.* Natural Language Understanding), до якої зокрема належать:

- семантичний поділ;
- узагальнення;
- виділення семантично значимих слів;
- перевірка правопису;
- пошук пов’язаних між собою документів та слів;
- побудова діалогових систем (у т.ч. “чат-ботів”);
- визначення інформаційного забарвлення тексту;
- пошук синонімів та омонімів.

Результати роботи NLP та NLU систем використовуються при побудові різноманітних інформаційно-аналітичних систем. Вхідні та проміжні дані у таких системах можна охарактеризувати як «великі дані» (Big Data)[2], за критеріями «три v»: їх обсягом (*англ.* volume), швидкістю появи та опрацювання (*англ.* velocity) та різноманітністю (*англ.* variety). Тому важливим фактором при опрацюванні текстових даних великих обсягів є не лише точність, але й ефективність. Можна стверджувати, що від функційних можливостей, якісних та кількісних характеристик програмних бібліотек NLU та NLP, які здійснюють попереднє опрацювання текстових даних, залежить точність та швидкість роботи таких аналітичних систем загалом. Зазначені особливості накладають додаткові обмеження на можливість застосування тієї чи іншої програмної бібліотеки NLP. Найпоширенішою програмною екосистемою та мовою програмування при побудові систем опрацювання великих даних є Java. Також, доволі часто використовуються, Python та Scala. Тому порівняльний аналіз бібліотек для

опрацювання великих даних варто здійснювати з урахуванням цих особливостей. Можливим, також є використання API хмарних сервісів [3] для задач NLP та NLU, проте, варто враховувати особливості їх тарифікації. Варто зазначити, що є значна кількість безкоштовних бібліотек та бібліотек з відкритим вихідним кодом, які надають цілу низку переваг розробнику, власнику та кінцевому споживачу таких систем. Авторами враховано ще один ключовий фактор при виборі таких бібліотек, а саме - підтримку української мови.

Таблиця 1. Бібліотеки опрацювання природної мови людини

Назва	Ліцензія	Підтримувана мова програмування	API для інших мов програмування	Мови що підтримуютья
OpenNLP	Apache Lic.v2.0	Java	Java	english
Stanford NLP	GNU General Public Licens	Java	Python (or Jython), Ruby, Perl, Javascript, F#, and other .NET and JVM language	english, germany, french, italian, ukrainian, russian
LingPipe	Alias-i	Java	Java	english, germany, french, italian
GATE	GNU General Public License	Java	Json, java	english, germany, french, italian, russian
LanguageTool	LGPL 2.1	Java	Java	english, germany, french, italian, ukrainian, russian
NLTK	Apache 2.0	Python	Python	english, germany, french, italian, ukrainian, russian
FreeLing	Affero GPL	Lex, C++, C	Python	english, germany, french, italian, russian
Apache Solr	Apache 2.0	Java	Java, JavaScript, Python, Ruby, JSON	english, germany, french, italian, ukrainian, russian
GoogleCloud Natural Language API	Apache 2.0		Java, JavaScript, Python, Ruby, C++, C#, C	english, germany, french, italian, russian

Таким чином серед усіх проаналізованих бібліотек найбільший функціонал та швидкодію має Stanford Core NLP[], проте найбільша кількість підтримуваних мов є у бібліотеці LanguageTool. Дані бібліотеки забезпечують поділ на речення та слова, POS-тегування, лематизацію. Нижча продуктивність LanguageTool при лематизації зумовлена використанням словників.

Література

1. Nitin Indurkhya, Fred J. Damerau Handbook Of Natural Language Processing Second Edition/ F. J. Damerau// Taylor and Francis Group, USA. 2010, 676

2. The App Solutions. Natural Language Processing Tools And Libraries [Електронний ресурс] Режим доступу: URL: https://theappsolutions.com/blog/development/nlp-tools/#contents_1

3. Stanford Core NLP [Електронний ресурс] Режим доступу: URL: https://nlp.stanford.edu/software/

4. LanguageTool [Електронний ресурс] Режим доступу: URL: http://wiki.languagetool.org/java-api