Метод визначення ключових слів англомовного тексту на основі інструментальних засобів пакету DKPro Core

Керівник: д.т.н., проф., Бісікало О. В.

Виконав: Яхимович О. В. 1КСУА-14 м

Визначення ключових слів

- **Актуальність дослідження**. Завдання виділення ключових слів з тексту виникає у бібліотечній справі, лексикографії та термінознавстві, а також в задачах інформаційного пошуку. В даний час обсяги і динаміка інформації, яка підлягає обробці в цих областях, роблять особливо актуальною задачу автоматичного визначення ключових слів, які можуть використовуватися для створення і розвитку термінологічних ресурсів, а також для ефективної обробки документів: індексування, реферування, кластеризації та класифікації.
- **Мета і завдання дослідження.** Мета роботи полягає у підвищенні точності визначення ключових слів з англомовного тексту.
- Для досягнення поставленої мети необхідно розв'язати наступні задачі: аналіз й порівняльна характеристика відомих алгоритмів та методів визначення ключових слів; інформаційна оцінка парсерингу тексту для задачі визначення ключових слів; розробка алгоритму знаходження ключових слів; реалізація алгоритм у середовищі обраного сучасного лінгвістичного пакету; провести експериментальну оцінку точності та повноти знаходження ключових слів.
- **Об'єкт дослідження** – процес обробки вербальної інформації для визначення ключових слів в тексті.
- **Предмет дослідження** – методи знаходження ключових слів в тексті.
- **Наукова новизна одержаних результатів.** Запропоновано новий метод визначення ключових слів, який, на відміну від існуючих, базується на знаходженні синтаксичних зв'язків між словоформами у реченнях англомовного тексту за допомогою інструментальних можливостей пакету DKPro Core. Запропонований метод дає змогу підвищити кількісні характеристики релевантності отриманих ключових слів, а саме повноту (за Жаккаром і абсолютну) і точність (за евклідовою і манхеттенською відстанями). Набула подальшого розвитку інформаційна модель оброблення тексту, яка, на відміну від існуючих, враховує додаткову інформацію процесів парсерингу речень, що дозволило уточнити чисельні оцінки для визначення ключових слів тексту.
- **Практичне значення одержаних результатів** роботи полягає у формальному описі методики знаходження ключових слів, створенні алгоритму її реалізації та розробці програми, що знаходить ключові слова на основі врахування зв'язків між словоформами у реченнях англомовного тексту.
- Створені моделі, алгоритми та програмні засоби можуть буди використані при вирішенні практичних задач комп'ютерної лінгвістики, які потребують знаходження ключових слів, наприклад, для підвищення точності аналізу контенту сайту і підняття позиції сайту в результатах пошуку.

Визначення ключових слів

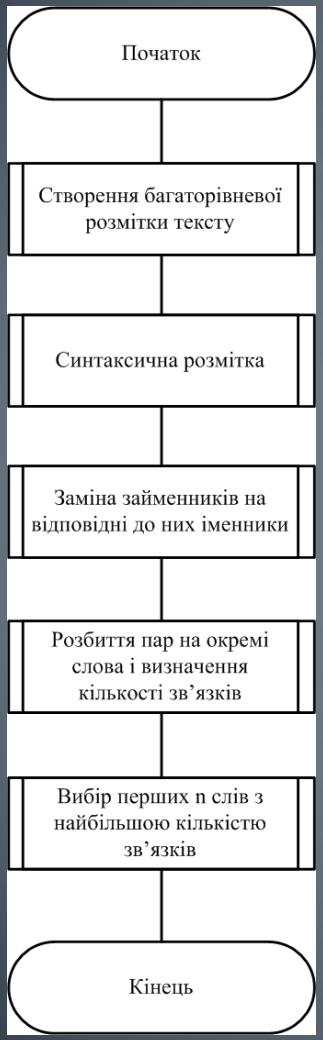
Розглянемо задачу визначення ключових слів тексту як певну інформаційну технологію, що має на вході текст, а на виході – множину з *l* ключових слів $W^k = \{w_1^k, \ldots, w_l^k\}$. Без применшення загальності будемо вважати, що текст *T* складається з *m* різних слів, а в окреме його *j*-те речення з *k* налічує *n* слів з *m* можливих, причому *m* >>*n* та *m* >> *l*. Більшість відомих методів визначення ключових слів тексту беруть за основу частотний словник тексту, який фактично є списком або упорядкованою множиною пар $D = \{< w_i, \ f_i >\}, \ i = \overline{1,m}$, де w_i – одне слово з *m*, а f_i – його частота ($f_i \geq f_{i+1}$, $i = \overline{1,m-1}$), що визначена для *T*. За певною фільтрацією окремих незначущих категорій слів ключовими вважають перші *l* слів зі списку *D*, тобто, дещо спрощено маємо $W^k = \{w_1, \ldots, w_l\}$.

Визначення ключових слів

Аналіз збільшення частоти значимих слів унаслідок урахування парних залежностей для різних типів речення

№ з/п	Склад речення / кількість слів	Тип речення та граф його дерева залежностей	Частотна формула	Кінце-ва частота
1.	Ab / 2	Словосполучення (Коріння дерева)	A+b	2
2.	Abc / 3	Лінійна трійка (Бережи скарби природи)	A+2b+c	4
3.	Abcd / 4	Лінійна четвірка (Отримав переклад слова дивного)	A+2b+2c+d	6
4.	Abcde / 5	Розгалудження (Густий ліс нізвідки завершився проваллям)	A+2b+c+3d+e	8
5.	Abcdef / 6	Група підмета (Сині примружені очі коханого говорили багато)	A+b+4c+d+2e+f	10
6.	Abcdef / 6	Група присудка (Досвідчений кінь борозну швидко відчує нюхом)	A+2b+c+d+4e+f	10
7.	Abcdef/6	Обидві групи (Старий дід Еол зобрав всіх вітрів)	A+3b+c+2d+e+2f	10

Визначення ключових слів



- DKPro Core – це набір програмних компонентів для обробки природної мови, що базується на Apache UIMA framework. Він був побудований з метою підвищення продуктивності дослідників, які працюють з автоматичним аналізом мови. Підхід DKPro Core полягає в тому, що дослідники повинні мати можливість зосередитися на своїх реальних наукових питаннях, а не на розробці технологій.
- Визначення ключових слів відбувається за кількома етапами:
- а) створення багаторівневої розмітки тексту;
- б) синтаксична розмітка, що враховує складні залежності між парами лем;
- в) заміна займенників в отриманих парах на відповідні до них іменники;
- г) розбиття пар на окремі слова і визначення кількості зв'язків;
- д) вибір перших n слів з найбільшою кількістю зв'язків, де n – кількість потрібних ключових слів.

Результати пошуку ключових слів

Для проведення експерименту використовувався текст статті з 1460 слів «A new pattern for historical geography: working with enthusiast communities and public history».
Ключові слова, задані автором: Participation, Public history, Enthusiast communities, Museums, Heritage.

Слова задані автором		власна розробка		rise-top		advego		seotool	
1	Participation		work		historical		historical		historical
2	Public	5	community	4	enthusiast	4	enthusiast	4	enthusiast
3	history		geography	5	communities		for	5	communities
4	Enthusiast	1	participation	1	participation	5	community	1	participation
5	communities	4	enthusiast		geography		this		work
6	Museums		geographer		work	6	museum		geography
7	Heritage	6	museum		research		geography		new

Кількісні характеристики

Повнота за Жаккаром, в даному випадку, це частка від ділення кількості знайдених ключових слів на різницю кількості можливих ключових слів заданих автором і знайдених програмно (в даному випадку по 7) і кількості знайдених ключових слів.

Абсолютна повнота знаходиться як відношення кількості правильно знайдених ключових слів до кількості ключових слів.

Евклідова відстань визначається за формулою:

$$d_e = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

n – кількість ключових слів.

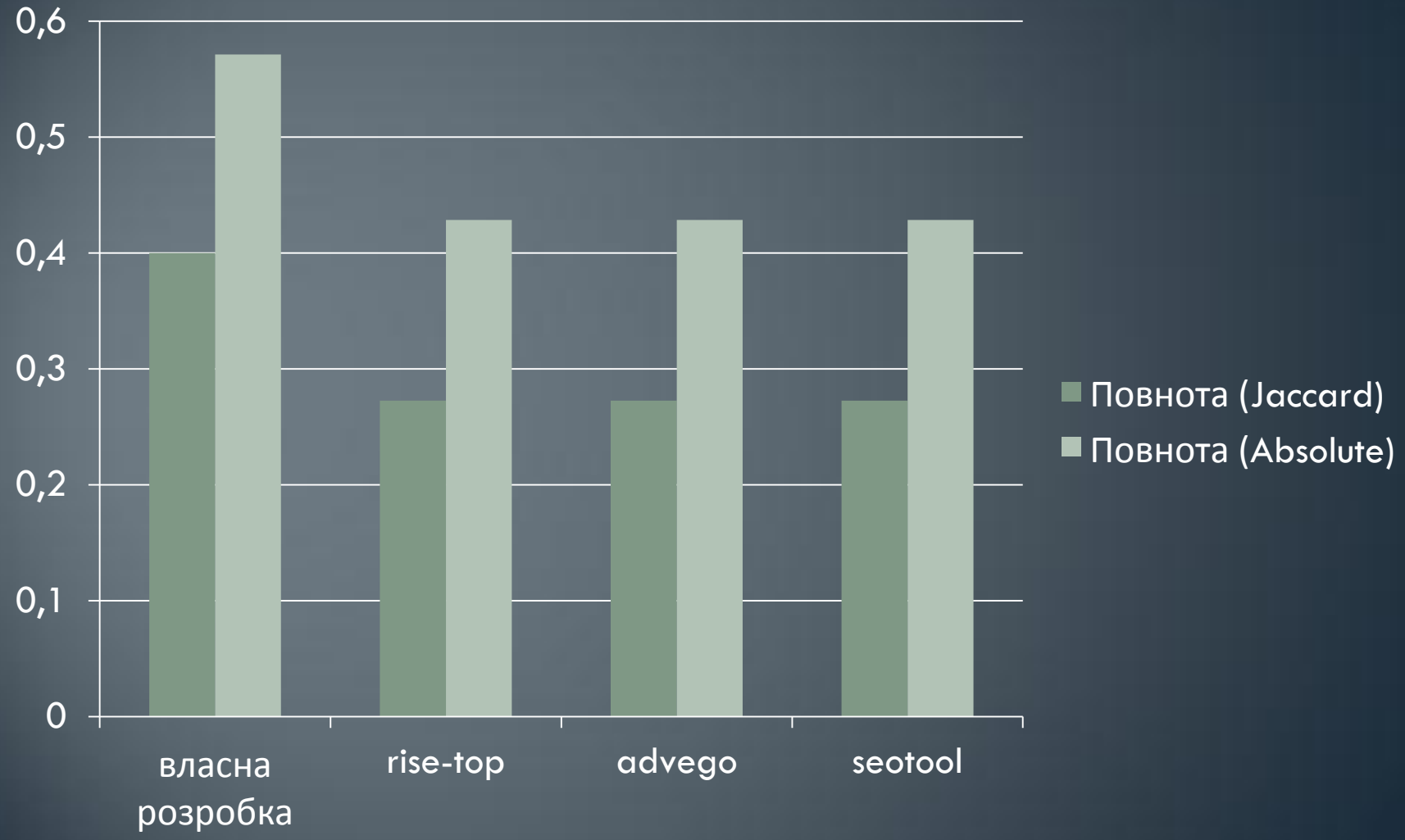
x_i – позиція i-го ключового слова визначеного автором.

y_i – позиція i-го ключового слова визначеного програмно.

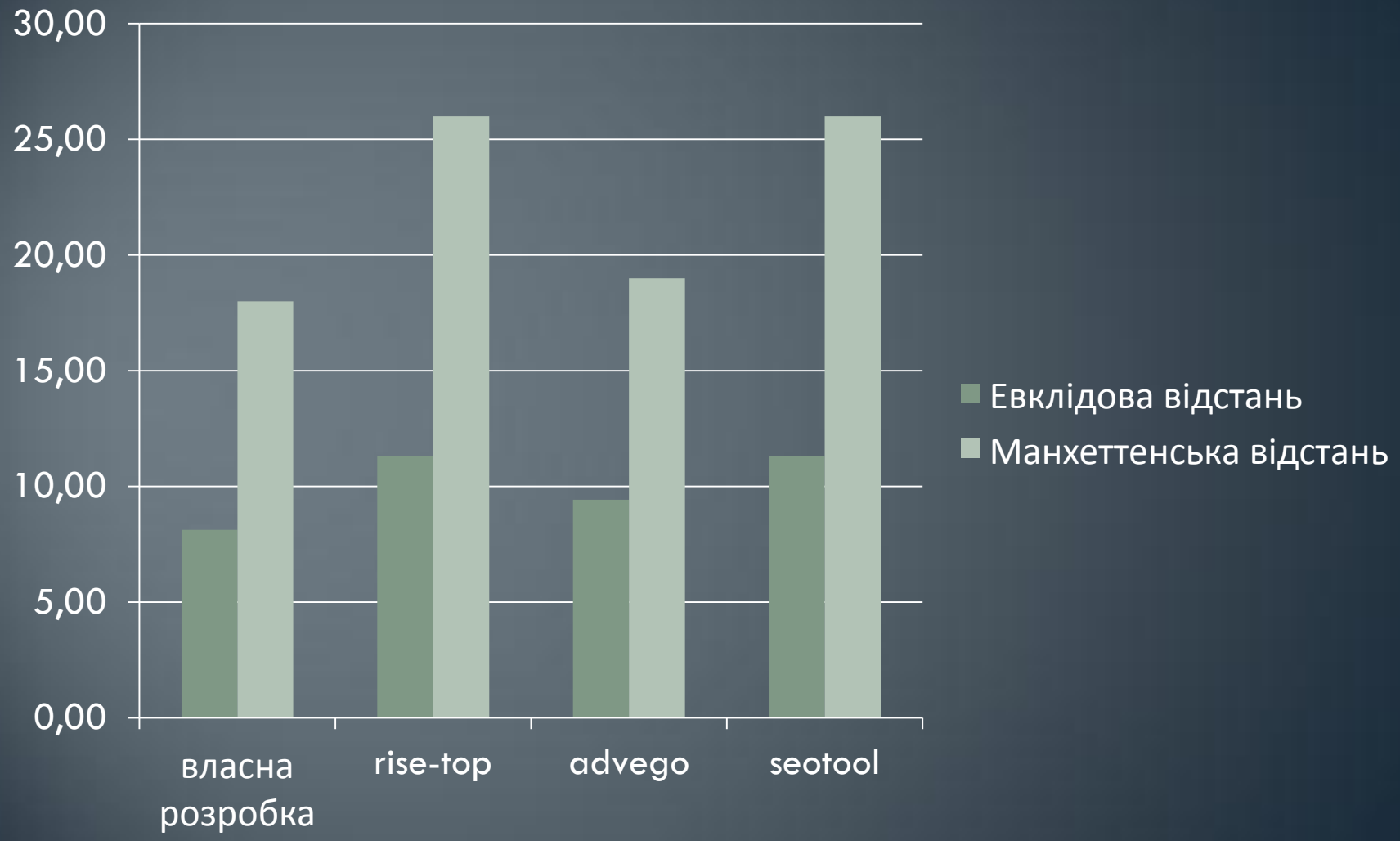
Манхеттенська відстань визначається за формулою:

$$d_m = \sum_{i=1}^{n} |x_i - y_i|$$

Гістограма повноти за Жаккаром і абсолютної



Гістограма евклідової та манхеттенської відстані



Публікації:

- **Конференції**
- Бісікало О.В. Моделювання процесів побудови парадигматичних зв’язків між словоформами на основі вимірювання текстової інформації / О.В. Бісікало, С.С. Траченко, О.В. Яхимович, А.І. Лісовенко // Вимірювання, контроль та діагностика в технічних системах (ВКДТС-2015): збірник тез доповідей III міжнар. наук. конф. (Вінниця, 27-29 жовтня 2015 р.). – Вінниця: ПП «ТД «Едельвейс і К», 2015. – С. 119-121.
- Бісікало, О. В; Яхимович, О. В. Автоматизація побудови тезарусу. XLIII регіональна науково-технічна конференція професорсько-викладацького складу, співробітників та студентів університету з участю працівників науково-дослідних організацій та інженерно-технічних працівників підприємств м. Вінниці та області відбулась 6-7 березня 2014 року.
- Бісікало, О. В; Яхимович, О. В. Застосування інструментальних засобів пакету DKPRO CORE для визначення ключових слів англомовного тексту. XLIV регіональна науково-технічна конференція професорсько-викладацького складу, співробітників та студентів університету з участю працівників науково-дослідних організацій та інженерно-технічних працівників підприємств м. Вінниці та області відбулась 3-6 березня 2015 року.
- Бісікало, О. В; Яхимович, О. В. Визначення ключових слів англомовного тексту з використанням технології DKPRO CORE. ВНТУ, Міжнародна Інтернет-конференція «Молодь в технічних науках: дослідження, проблеми, перспективи (МНТ-2015)», 23-26 квітня 2015 р. С. 72-74.
- Бісікало, О. В., Лісовенко, А. І., Яхимович, О. В., Траченко, С. С. Підтримка діалогу з навчальним контентом. Міжнародна конференція з адаптивних технологій управління навчанням ATL-2015, 23-25 вересня 2015 р. С. 97-100.
- **Статті**
- Бісікало, О. В; Яхимович, О. В. Метод визначення ключових слів англомовного тексту на основі Dkpro Core. Технологический аудит и резервы производства, 2015, № 2 (21), том 1. С. 26-30.
- Бісікало, О. В., Лісовенко, А. І., Яхимович, О. В., Траченко, С. С. (2015). Визначення змістовних ознак тексту на основі аналізу зв’язків між лексичними одиницями. Вісник Національного технічного університету «ХПІ», 2015, № 21(1130). С 83-89.
- Бісікало, О. В; Яхимович, О. В. Знаходження ключових слів англомовного тексту за допомогою інструментальних засобів пакету DKProCore. Інформаційні технології та комп’ютерна інженерія, 2015, №2. С. 10-14.

Висновки

- Оскільки краща якість обробки тексту досягається лінгвістичними методами або ж при їх комбінації зі статистичними, систему автоматичного визначення ключових фраз з тексту природною мовою слід розробляти з використанням морфологічного словника (лексикону) і синтаксичних правил. Ці дані визначаються попередньо і зберігаються в базі даних. Текст підлягає обробці аналізатором, який виробляє інформацію про розділення тексту на абзаци, речення та окремі слова, що необхідно для подальшого оброблення. Кожне слово, виділене аналізатором, піддається морфологічному аналізу з метою побудови морфологічної інтерпретації, визначення основи слова і формування леми. На основі наявної інтерпретації тексту виконується побудова та наповнення синтаксичних груп і виявлення відношень між ними.
- В роботі запропоновано метод визначення ключових слів, що базується на використанні додаткової інформації про складні залежності між членами англомовного речення. Для функціональної реалізації аналізатора тексту обрано популярний лінгвістичний пакет DKPro Core. Проведені експериментальні дослідження теоретичного обґрунтування методу підтвердили його якісні та кількісні переваги у порівнянні з відомими аналогами. Для англомовного тексту обсягом 1460 слів отримано збільшення повноти визначення ключових слів (на 31,8% за Жаккаром та на 25% за абсолютним значенням) і покращення точності (на 14% за евклідовою і на 5,3% манхеттенською відстанями).
- Якість отриманих результатів потенційно можна підвищити через окремий аналіз частин мови, оскільки ймовірність релевантності ключового слова, наприклад, іменника і прислівника буде відрізнятися. Окрім цього, варто оцінити збільшення частотних показників для ключових слів шляхом реалізації наявних в DKPro Core компонентів для визначення кореференційних зв’язків.

Дякую за увагу