НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ імені
ІГОРЯ СІКОРСЬКОГО»
Інститут прикладного системного аналізу
Кафедра математичних методів системного аналізу

Магістерська дисертація

на тему: “Система аналізу неструктурованих текстових даних”

Виконав:
Студент групи КА-61м
Шибирин Ігор Олегович

Науковий керівник:
ст.н.с. ІПС НАН України,
к.ф.-м.н, ст.н.с.
Ігнатенко О.П.

Київ - 2018

Актуальність роботи

- ✓ Велика кількість інформації
- ✓ Конкурентна боротьба компаній
- ✓ Підтримка прийняття бізнес-рішень

Підзадачі:

- ✓ Побудова алгоритму визначення аспектів
- ✓ Побудова алгоритму визначення настрою

Мета роботи:

Розробка та дослідження автоматичної системи визначення настроїв щодо аспектів продукту, що дає можливість усувати інформаційну невизначеність при прийнятті управлінських рішень

Об’єкт дослідження:

Складні системи аналізу неструктурованих текстових даних.

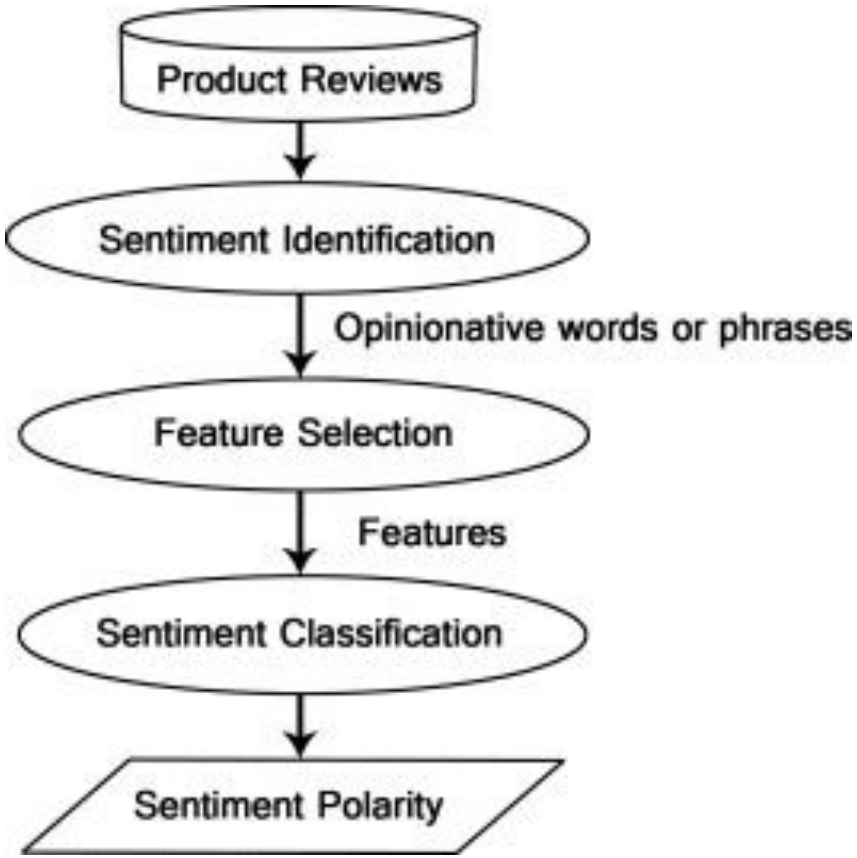
Предмет дослідження:

Методи та системи аналізу неструктурованих текстових даних.

Постановка задачі

- ✓ Розробити систему аналізу неструктурованих текстових даних на основі вже існуючих систем.
- ✓ Розробити програмне забезпечення для аналізу. Програмне забезпечення має допомогти аналітику приймати рішення для виконання поставленої задачі на основі оброблених текстових даних.
- ✓ Розробити інтерфейс для перегляду та подальшого використання отриманих результатів.

Класична схема визначення настрою



Складові висловлення

- Сутність
- Аспект
- Орієнтація настрою
- Власник думки
- Час

Додаткові поняття

- **Вираз сутності**: відповідає дійсному слову або фразі, написаній користувачем для позначення або позначення сутності. У результаті об'єкти є узагальненням кожного виразу сутності, використаного в аналізованих документах, або конкретної реалізації виразу сутності;
- **Вираз аспекту:** як і для виразу суб'єкта, аспектним виразом є фактичне слово або фраза, написані користувачем для позначення аспекту. Таким чином, аспекти також є загальними поняттями, які містять кожне вираження аспекту.

Модель сутності

Сутність e_i відображається сам по собі як ціле а також як скінченний набір аспектів, $A_i = \{a_{i1}, a_{i2}, ..., a_{in}\}$. Сутність може бути виражена будь-яким з кінцевих наборів виразів сутності: $EE_i = \{ee_{i1}, ee_{i2}, ..., ee_{is}\}$. Кожен аспект a_{ij} A_i сутності може бути виражений будь-яким з кінцевих наборів виразів аспекту $AE_{ij} = \{ae_{ij1}, ae_{ij2}, ..., ae_{ijm}\}$ З іншого боку, документ думки $d_k \in D$ містить думки про сукупність сутностей $e_1, e_2, ..., e_r$ з набору власників думок $h_1, h_2, ..., h_p$ Погляди на кожну сутність e_i виражаються на самій сутності а також на підмножині її аспектів A_{ik}.

Алгоритм визначення орієнтації слова

Algorithm 1 Word Orientation

```
1: if word is in opinion_words then
2:     mark(word)
3:     orientation ← Apply Opinion Word Rule(marked_word)
4: else
5:     if word is in neutral_words then
6:         mark(word)
7:         orientation ← 0
8:     end if
9: end if
10: if word is near a too_word then
11:     orientation ← Apply Too Rules(orientation)
12: end if
13: if word is near a negation_word then
14:     orientation ← Apply Negation Rules(orientation)
15: end if
16: return orientation
```

Правила

- Правило слова
- Правило заперечення
- Правило надмірності

Орієнтація думки

Algorithm 2 Opinion Orientation

```
1: if but_word is in sentence then
2:     orientation ← Opinion Orientation(aspect,marked_words,but_clause)
3:     if orientation ≠ 0 then
4:         return orientation
5:     else
6:         orientation ← Opinion Orientation(aspect,marked_words,not but_clause)
7:         if orientation ≠ 0 then
8:             return -1 × orientation
9:         else
10:            return 0
11:        end if
12:    end if
13: else
14:    for all aspect_position in aspect do
15:        for all aspect_word in aspect_position do
16:            for all word in marked_words do
17:                suborientation += Word Orientation(word) / WD(aspect_word,word)
18:            end for
19:            orientation += suborientation
20:        end for
21:        final_orientation += orientation
22:    end for
23:    if final_orientation > 0 then
24:        return 1
25:    else
26:        if final_orientation < 0 then
27:            return -1
28:        else
29:            return 0
30:        end if
31:    end if
32: end if
```

Правило агрегування аспектних слів

$$score(aw_{ij}, s) = \sum_{ow_j \in s} \frac{score(ow_j)}{WD(ow_j, aw_{ij})}$$

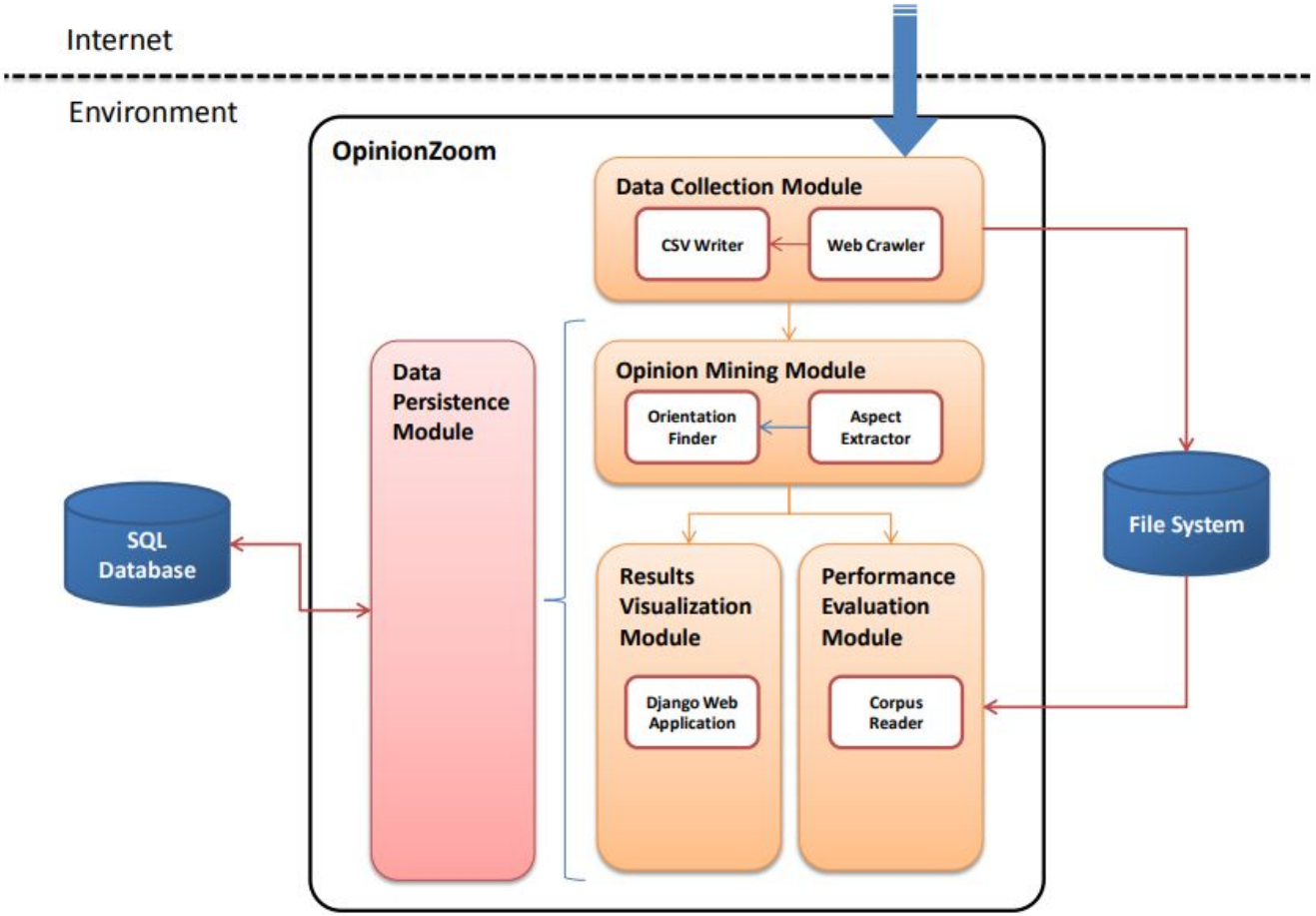
Правило агрегування аспектів

$$score(a_i, s) = \sum_{aw_{ij} \in AW_i} score(aw_{ij}, s)$$

Правило агрегації позиції

$$fscore(a_i, s) = \sum_{k=1}^{t} score(a_i^k, s)$$

Архітектура системи



Архітектура системи

Модуль збору даних (DCM)

Модуль видобутку думки (OMM)

Модуль візуалізації результатів (RVM)

Модуль оцінювання продуктивності (PEM)

Деталі відгуків

Кількість відгуків	100
Загальна кількість речень	487
Кількість речень з думками	376
Загальна кількість/кількість з думками	77,21%

Перевірка результатів

Тип аспекту	Відгуки	
	Кількість	Відношення
Явний	229	73,87%
Явний і неявний	30	9,68%
Неявний	51	16,45%
Загальна кількість	310	100%

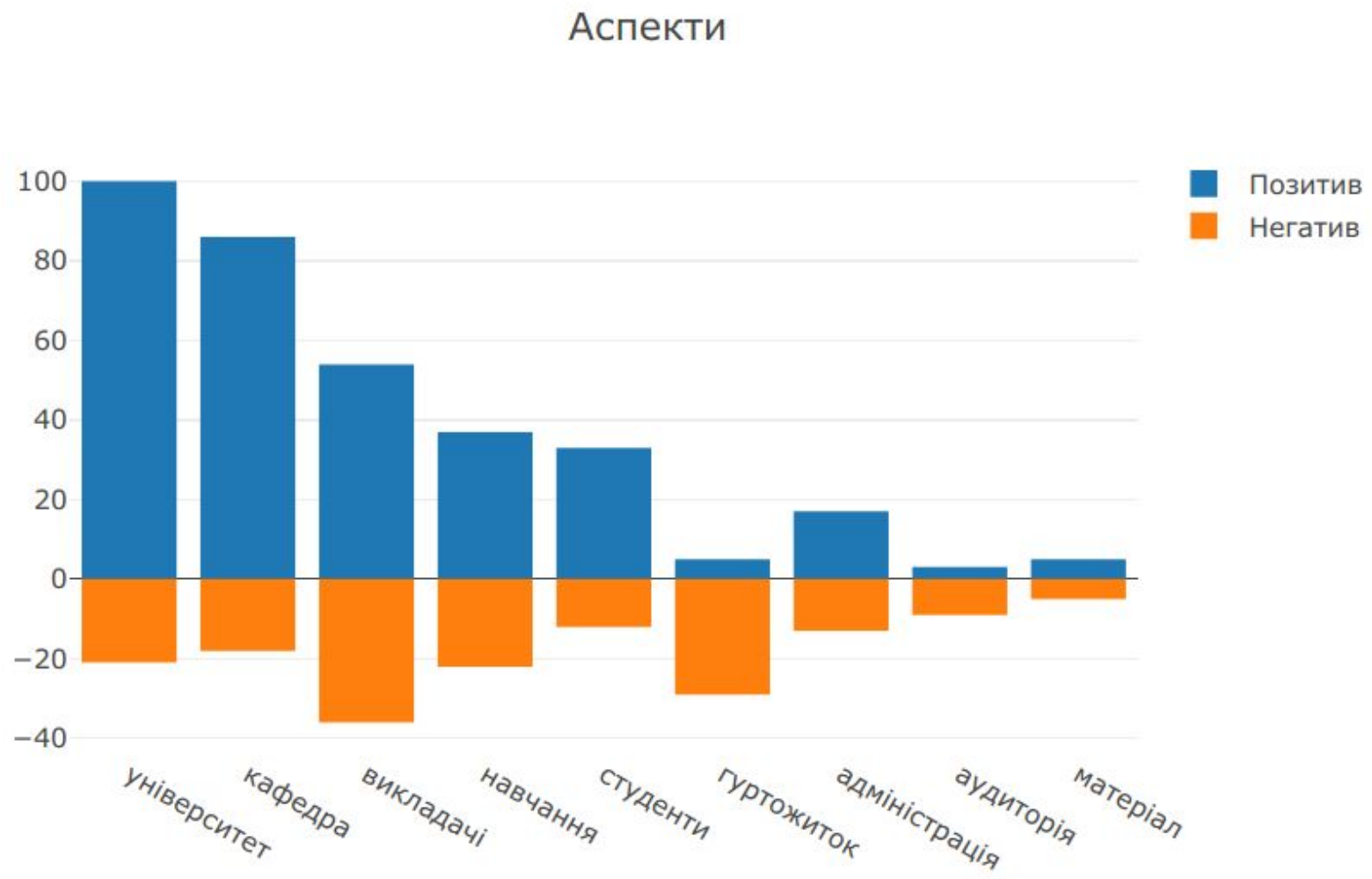
Оптимальні значення

Індекс	P	R	F-m
Явно отримані аспекти	33%	29%	31%
Суб’єктивна класифікація	79%	93%	85%
Класифікація настрою	89%	93%	91%

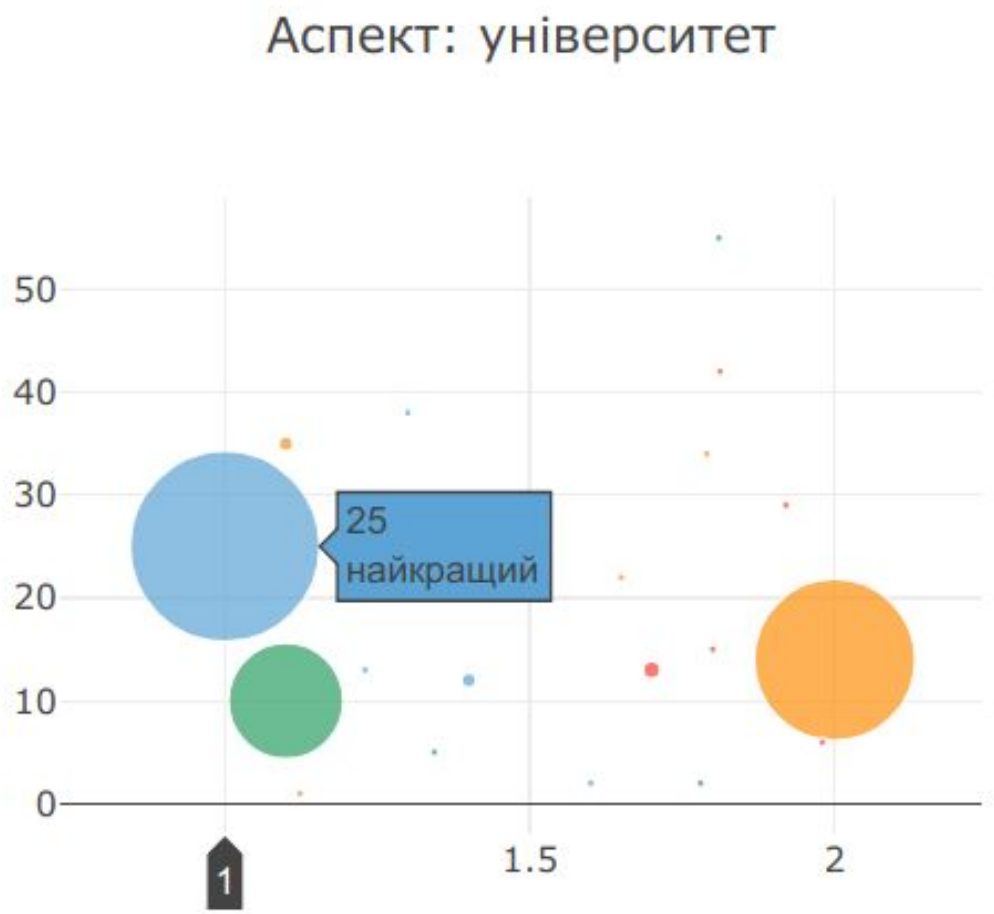
Порівняння з підходом Лю

Індекс	Точність (P)		Відклик (R)		F-міра	
	тут	Б. Лю	тут	Б. Лю	тут	Б. Лю
Явно отримані аспекти	33%	69%	29%	59%	31%	64%
Суб’єктивна класифікація	79%	63%	93%	71%	85%	67%
Класифікація настрою	89%	90%	93%	90%	91%	90%

Візуалізація результатів



Візуалізація результатів



Висновки

Для реалізації поставленої задачі були виконані наступні кроки:

- досліджено та проаналізовано існуючі методи аналізу неструктурованих текстових даних;
- проаналізовано алгоритми та підходи до визначення настрою тексту, виявлено їх переваги та недоліки;
- проаналізовано алгоритми та підходи до визначення аспектів продукту;
- описано загальну архітектуру системи аналізу неструктурованих текстових даних;
- розроблено методу визначення аспектів та відношення автора до них;
- розробно візуального представлення результатів роботи методу;
- проведений аналіз стартап потенціалу розробленого методу.

Дякую за увагу!