Інтелектуальний аналіз текстів на основі самоорганізації

Студент групи КА-34
Полубок Артем Михайлович
Керівник роботи:
д.т.н. проф. Данилов В. Я.

Об'єкт дослідження

Набір відформатованих текстів, написаних українською мовою.

Предмет дослідження

Методи морфологічного аналізу та синтаксичного аналізу, метод класифікації на основі нейронної мережі.

Актуальність задачі

Необхідність створення системи обробки природньої мови для української мови. Для цього необхідно створити систему, яка буде реалізовувати алгоритм стемінгу української мови та вилучення ключових слів, що надасть можливість автоматизувати велику кількість процесів пов'язаних з обробкою україномовних текстів.

Постановка задачі

- Ознайомитися з існуючими системами та методами, які вирішують поставлену задачу або її модифікацію;
- Запропонувати власний метод вилучення ключових слів для української мови;
- Реалізувати метод у вигляді програмного продукту;
- Проаналізувати результати роботи системи.

Що таке ключове слово?

Ключове слово в аналізі текстів – особливо важливі, загальнозрозумілі, змістовні і показові для окремо взятої культури слова або словосполучення в тексті, набір котрих здатний дати високорівневий опис змісту документу, забезпечивши компактне представлення та зберігання змісту документа в пам’яті.

Навіщо потрібне вилучення ключових словосполучень?

- Швидкий пошук документів за ключовими словами;
- Зв'язування різних документів між собою за ключовими словами;
- Кластерізація та каталогізація документів на основі ключових слів;
- Визначення тематики документа за ключовими словами.

Основні системи, які вирішують схожу задачу.

- Фреймворк Sztergak – розробники: Gabor Berend і Richard Farkas;
- Система Humb – розробники: Patrice Lopez і Laurent Romary;
- Система Wingnus – розробники: Thuy Dung Nguyen і Minh-Thang Luong;
- Система KP-Miner – розробники: Samhaa R. El-Beltagy і Ahmed Rafea.

Загальна схема вилучення ключових слів для даних систем

- Визначення набору ознак, які характеризують ключові слова;
- Обробка тексту, вилучення зайвих елементів(пунктуації, незначущих слів);
- Відбір кандидатів;
- Розрахунок вектора значень атрибутів для кожного кандидата;
- Обробка отриманих даних;
- Відбір кандидатів.

Критерії оцінки системи

- Точність (Precision) - відношення числа релевантних ключових слів знайдених автоматично, до загальної кількості знайдених ключових слів в документі;
- Повнота (Recall) - це відношення числа релевантних ключових слів знайдених автоматично, до загальної кількості релевантних ключових слів в документі;
- F-міра (F-score, F-measure) - об'єднання точності і повноти в одній усередненій величині, визначається як зважене гармонійне середнє точності і повноти.

Для оцінки системи використовувався метод cross-validation по K блокам.

Результати роботи оглянутих систем на запропонованих розробниками даних

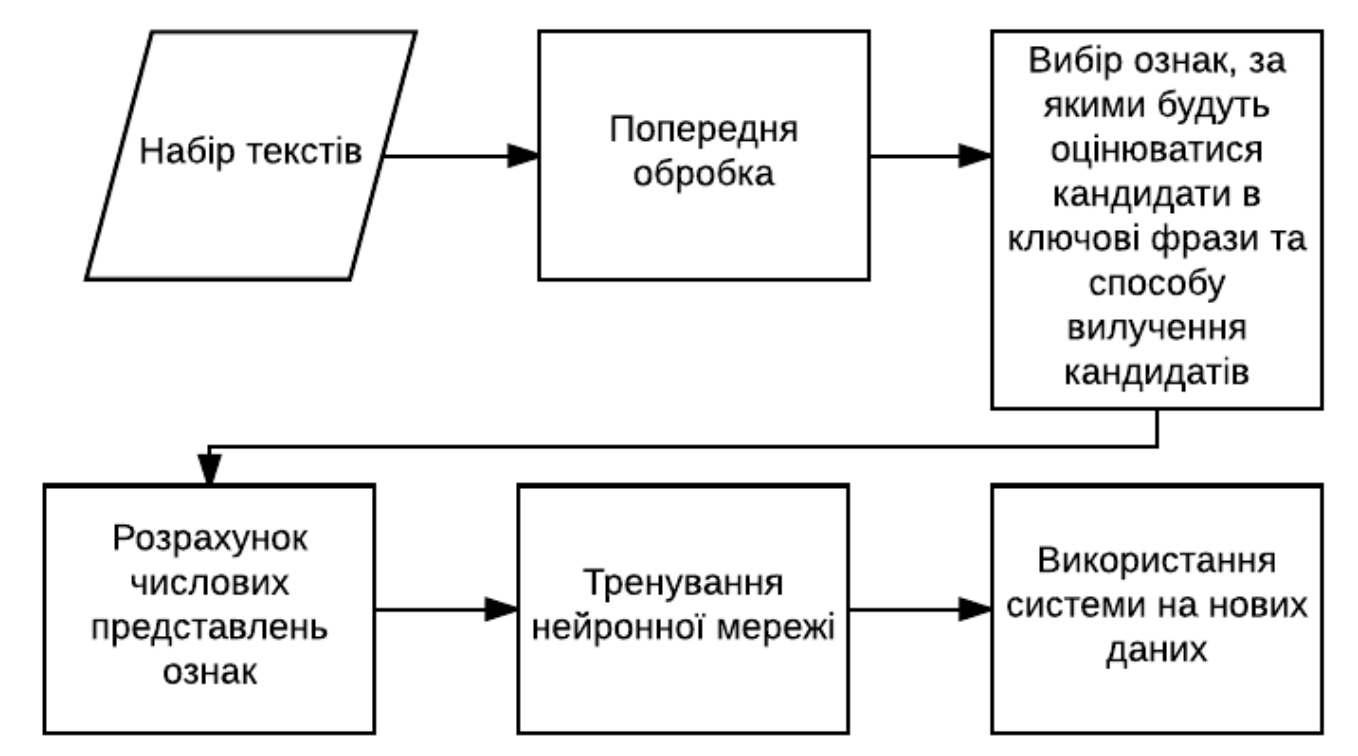
Ефективність розглянутих систем для топ-10 ключових фраз

Команда	Точність, %	Повнота, %	F-міра, %
Sztergak	37.80	25.78	30.65
Humb	32.00	21.83	25.95
Wingnus	30.50	20.80	24.73
KP-Miner	28.60	19.51	23.20

Ефективність розглянутих систем для топ-15 ключових фраз

Команда	Точність, %	Повнота, %	F-міра, %
Sztergak	30.80	31.51	31.15
Humb	27.20	27.83	27.51
Wingnus	24.93	25.51	25.22
KP-Miner	24.93	25.51	25.22

Схема запропонованого методу вилучення ключових слів



Попередня обробка тексту

- Вилучення пунктуації;
- Вилучення стоп-слів – слова, які виконують службову роль у тексті і не несуть змістовного навантаження;
- Створення словника словоформ, де ключ – це основа слова, отримана застосуванням алгоритму стемінгу, а значення – кортеж, який складається з усіх словоформ цієї основи, знайдених в тексті.

Запропоновані ознаки для оцінки кандидатів

- TF-IDF – ознака, заснована на частоті вживання фрази в тексті, при обробці документів однієї тематики, враховується також факт наявності фрази в кожному документі набору;
- Перше входження слова – позиція першого входження слова в тексті відносно загальної кількості слів в тексті;
- Коефіцієнт входження фрази, обчислюється коли фраза є підмножиною інших фраз;
- Довжина фрази кандидата;
- Наявність фрази в назві тексту – двійкова ознака.

Методи вилучення кандидатів з тексту

1. Після попередньої обробки беруться усі фрази довжиною не більше заданого значення;
2. Формуються набори слів, які не є розділені пунктуацією або стоп-словами, емпірично було встановлено, що довжина ключової фрази рідко перевищує 3 слова, що дозволило встановити константне значення для першого методу.

Призначення нейронної мережі

Задача нейронної мережі класифікувати кандидатів як позитивних (кандидати, які є ключовими фразами) та негативних (кандидати, які не є ключовими словами).

Після цього кандидати ранжуються відповідно до частоти та значення коефіцієнту першого входження(якщо ознака першого входження була включена).

Побудова нейронної мережі

Вибрана структура нейронної мережі – багатошаровий перцептрон з одним прихованим шаром. Кількість вхідних нейронів відповідає кількості обраних ознак, кількість нейронів прихованого шару була встановлена емпірично та дорівнює 165, вихідний шар має 1 нейрон, оскільки вирішується задача класифікації з двома класами.

Кількість нейронів	Значення F-міри
150	17.52 %
160	18.10 %
165	18.51 %
170	18.34 %
180	18.21 %

Градієнтний алгоритм навчання

Суть алгоритму: зведення задачі навчання до задачі локальної оптимізації. Критерій оптимізації – мінімізація функції помилки:

$$e(w) = \sum_{i=1}^{M} (d_i - y_i(w))^2 \to min,$$

Ваги навчаються наступним чином:

вхідні ваги: $W_{ij}^I(n+1) = W_{ij}^I(n) - \gamma_{n+1} \frac{de(W)}{dW_{ij}^I}$

вихідні ваги: $W_{ij}^0(n+1) = W_{ij}^0(n) - \gamma_{n+1} \frac{de(W)}{dW_{ij}^0},$

Недоліки алгоритму:

- повільно збігається.
- знаходить лише локальний екстремум.

Генетичний алгоритм навчання

Суть алгоритму: зведення задачі навчання до задачі глобальної оптимізації. Критерій оптимізації – мінімізація функції середньоквадратичної помилки: $E(W) = \frac{1}{M}\sum_{k=1}^{M}(d_k - y_k(W))^2 \to min$

Задається початковий набір ваг, далі обчислюється якість прогнозування: $FI(W_i) = C - E(W_i) \to max$

Схрещуються існуючі пари: $W_i(0) \oplus W_k(0) = W_i(1) + W_k(1)$

Далі відбувається генерація нащадків із застосуванням мутації: $w'_{ij} = w_{ij}(n) + \xi(n)$

Далі вибираємо нову популяція, шляхом вибору N кращих з батьків та нащадків за критерієм FI.

Недоліки:

- він потребує значних обчислювальних витрат;
- ряд параметрів визначається експериментально.

Навчання мережі

Були запропоновані два алгоритми навчання: градієнтний та генетичний алгоритм. Після порівняння виявилося, що генетичний алгоритм при значно більшій обчислювальній складності демонструє не суттєве покращення точності в порівнянні з градієнтним алгоритмом, тому для кінцевого програмного продукту, було обрано градієнтний алгоритм навчання.

Результати порівняння алгоритмів навчання

Номер експерименту та тестової вибірки.	Генетичний алгоритм(значення F-міри)	Градієнтний алгоритм(значення F-міри)
1	18.88 %	18.67 %
2	18.03 %	18.02 %
3	18.07 %	17.86 %
4	18.38 %	18.08 %
5	18.61 %	18.38 %
6	18.71 %	18.45 %
7	18.07 %	17.91 %
8	18.79 %	18.71 %
9	18.88 %	18.80 %
10	18.67 %	18.29 %

Результати роботи системи на запропонованому наборі текстів

Набір ознак	Середнє значення F-міри
Ч	11.87 %
Ч + ПВ	14.76 %
Ч + ПВ + ДФ	16.56 %
Ч + ПВ + КВ	16.32 %
Ч + ПВ + ДФ + КВ	18.12 %
Всі ознаки	18.51 %

Скорочення назв ознак:

Ч – частота

ПВ – перше входження

КВ – коефіцієнт входження

ДФ – довжина фрази

Приклад роботи системи

Дано текст, тематика – фізика. Текст розкриває такі поняття: ядерна та атомна фізика, явище фотоефекту. Назва тексту – Атомна і ядерна фізика.

Запропоновані ключові слова: атом, атомна фізика, випромінювання, електрон, енергія, фотоефект, фотон, ядерна реакція, ядерна фізика, ядро.

Вилучені ключові слова: атом, атомна фізика, випромінювання, енергія, нуклон, фізика, фотоефект, фотон, ядерна фізика, ядро.

Висновки

- Виконано огляд систем та алгоритмів, які вирішують дану задачу.
- Запропоновано алгоритм для вилучення ключових слів для української мови.
- Створено програмний продукт із застосуванням мови Python.
- Проведено тестування системи на тестовому наборі даних із застосування методу cross-validation.

Дякуємо за увагу