ДОСЛІДЖЕННЯ МУЛЬТИАГЕНТНИХ СИСТЕМ ДЛЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ТЕКСТОВИХ ДАНИХ

Автор роботи:
Студент групи ДА-62
Сиротюк Олександр Васильович

Об’єкт дослідження:

- Мультиагентні системи для аналізу великих масивів текстових даних.

Предмет дослідження:

- Використання мультиагентних систем для вирішення задачі побудови словнику предметної області на великих масивах текстових даних

Метою даної роботи є:

- Дослідження методів аналізу текстових даних для знаходження подібних слів та ключових слів предметної області.
- Дослідження архітектур мультиагентних систем для аналізу масивів текстових даних та розробка власної для доповнення словника предметної області

Результатом даної роботи є:

- Розроблена мультиагентна система для доповнення словників предметної області на основі масивів текстових даних

Актуальність даної роботи:

Актуальність роботи обумовлена стрімким зростанням кількості текстової інформації, котра генерується у світі, а також тенденцією до збільшення даного показнику.

Також, поява нових засобів знаходження семантичної близькості надає можливість використати їх до задач, котрі ще не вирішувалися іншими дослідниками.

Появ та популяризація архітектур на основі MAC для аналізу великих даних (Лямбда та Каппа, IoT архітектури)

МАС для аналізу текстових даних

Під поняттям МАС в аналізі текстових даних будемо вважати: самоорагізовану систему, котра виконує поставлену задачу базуючись на розподіленні завдань між групою агентів. Агенти можуть представляти собою:

- Раціональні агенти.
- Незалежні процеси.
- Потоки виконання.

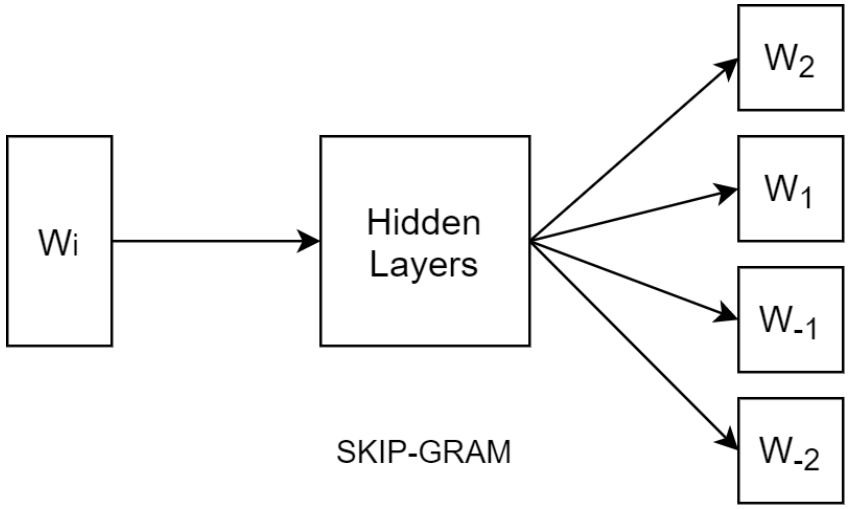
Задача та цільова архітектура МАС

На етапі обробки наукових робіт та книг, присвячених мультиагентних систем для аналізу текстових даних, була сформована задача по доповненню словників предметної області.

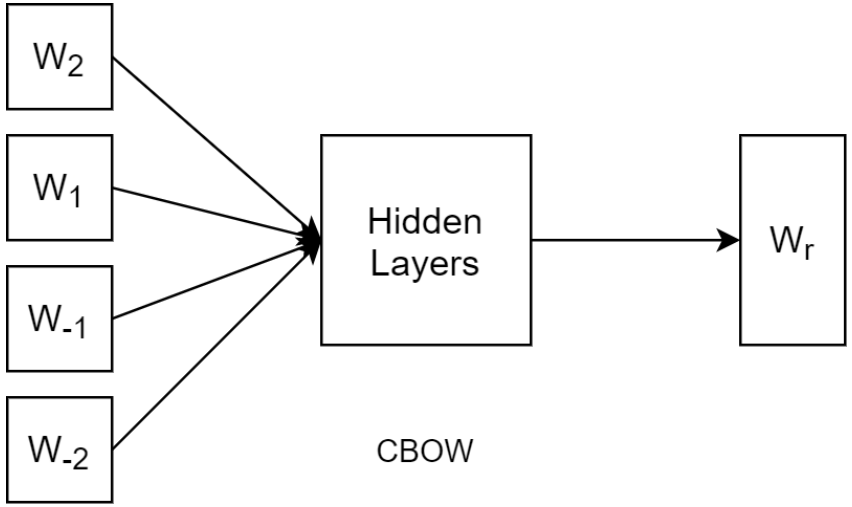
Методи для аналізу текстових даних

В даній роботі основною одиницею аналізу тексту є словосполучення. Для аналізу документів, або ж корпусу текстів основними варіантами виступають:

- Векторизація моделлю **Word2Vec**.
- Використання частотного аналізу **TF-IDF**.
- Використання алгоритму **TextRank.**
- Використання алгоритму **RAKE.**



SKIP-GRAM



CBOW

Word2Vec

Word2Vec – модель нейронної мережі, існує два види:

- Архітектура Skip-Gram – обрано для даного дослідження.
- Архітектура CBOW.

TF-IDF

TF-IDF – метод для частотного аналізу корпусу текстових даних та створення набору ключових слів, котрі описують корпус. Складається з двох основних частин:

- **TF** – частота терміну
- **IDF** – обернена частота документу

$$\boldsymbol{TF} = \frac{NGrammOccurence}{DocumentLen}$$

$$\boldsymbol{IDF} = log(\frac{DocumentsInCorpus}{DocumentsWithTermInside})$$

$$\boldsymbol{TF - IDF} = TF \cdot IDF$$

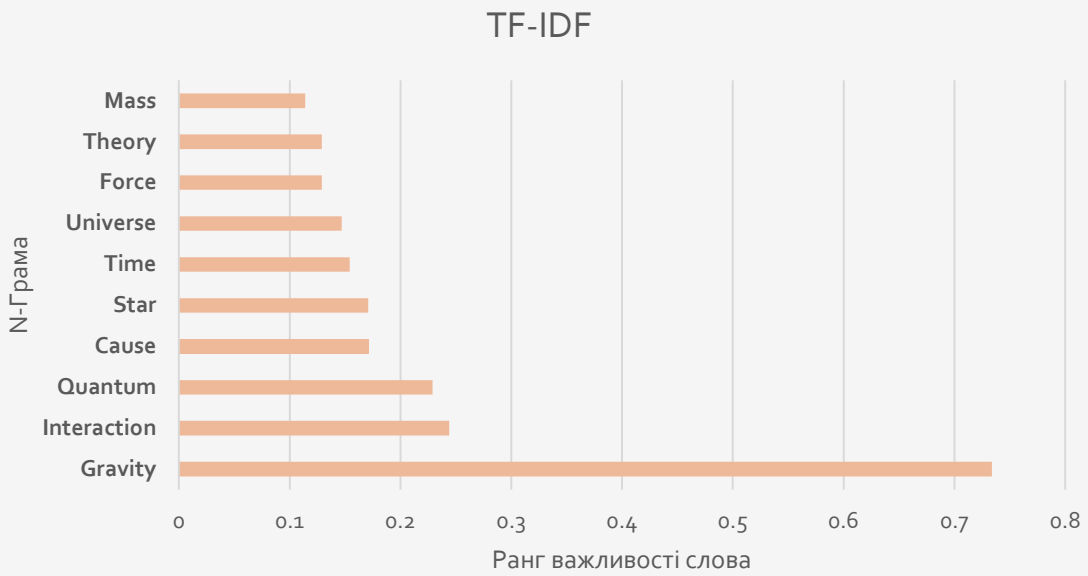
Підготовка даних

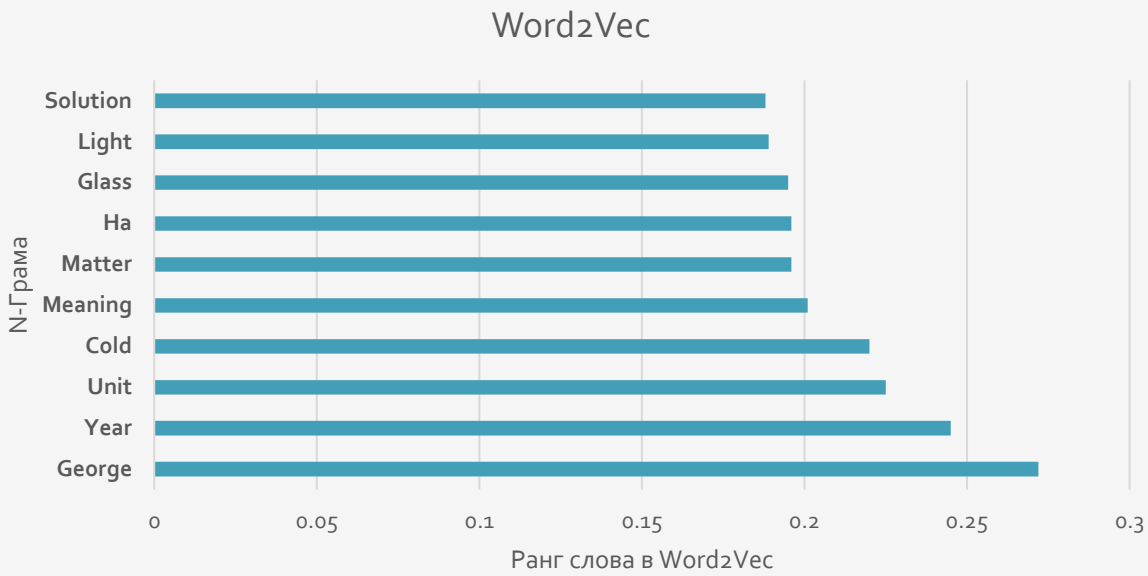
Перед аналізом даних потрібно провести процес нормалізації та індексації даних:

- Видалення знаків пунктуації.
- Приведення слів тексту до нижнього регістру.
- Розбиття слів на n-грами.
- Лематизація.
- Видалення слів, котрі не є іменниками.
- Індексація.

Даний процес є основним та використовується як при порівнянні методів, так і в розробленій моделі системи.

Результати TF-IDF та Word2Vec

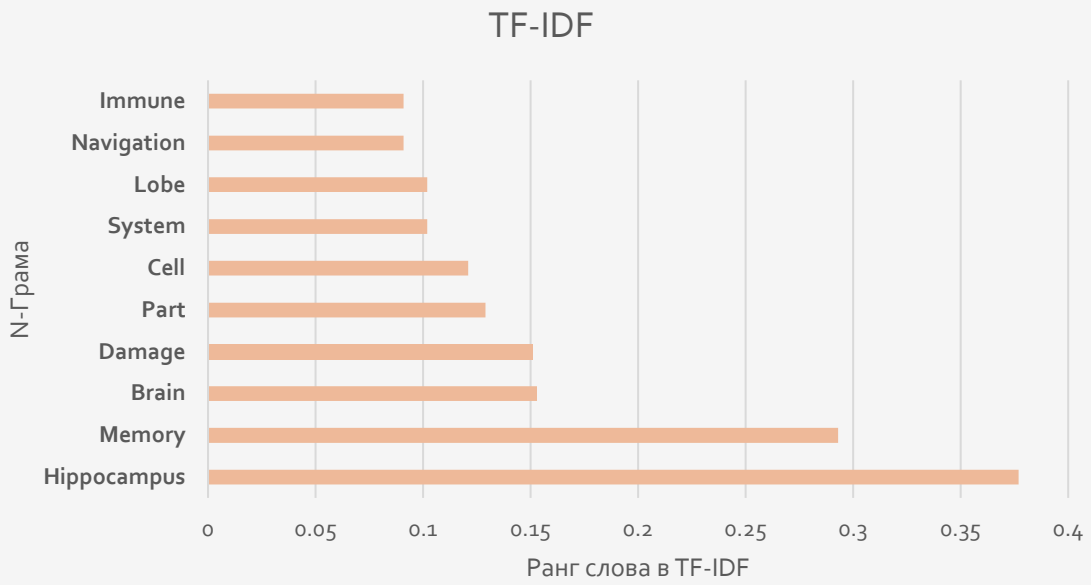


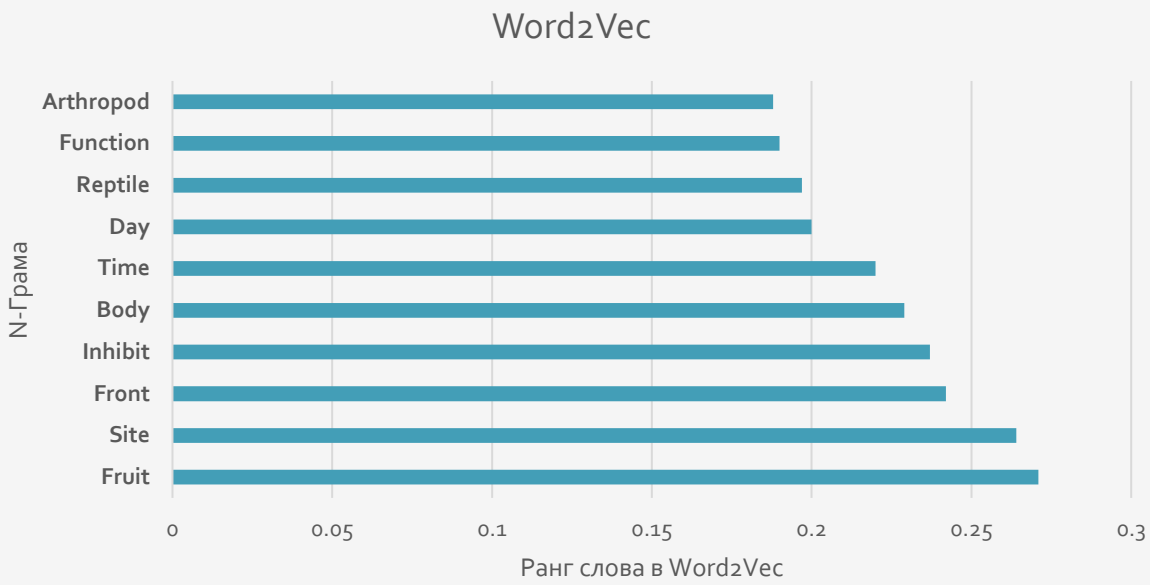


Характеристика	Значення
Кількість документів	11 документів
Розмір корпусу в символах	27701 символів
Розмір корпусу в словах	5732 слова
Розмір корпусу в байтах	27100 байт
Розмір словника в N-грамах	135

Предметна область - фізика

Результати TF-IDF та Word2Vec





Характеристика	Значення
Кількість документів	6 документів
Розмір корпусу в символах	43105 символів
Розмір корпусу в словах	12527 слова
Розмір корпусу в байтах	62311 байт
Розмір словника в N-грамах	44

Предметна область - біологія

RAKE та TextRank

- **RAKE** – алгоритм призначений для без контекстного витягу ключових фраз з документу, без урахування вживання ключових фраз у всьому корпусі.

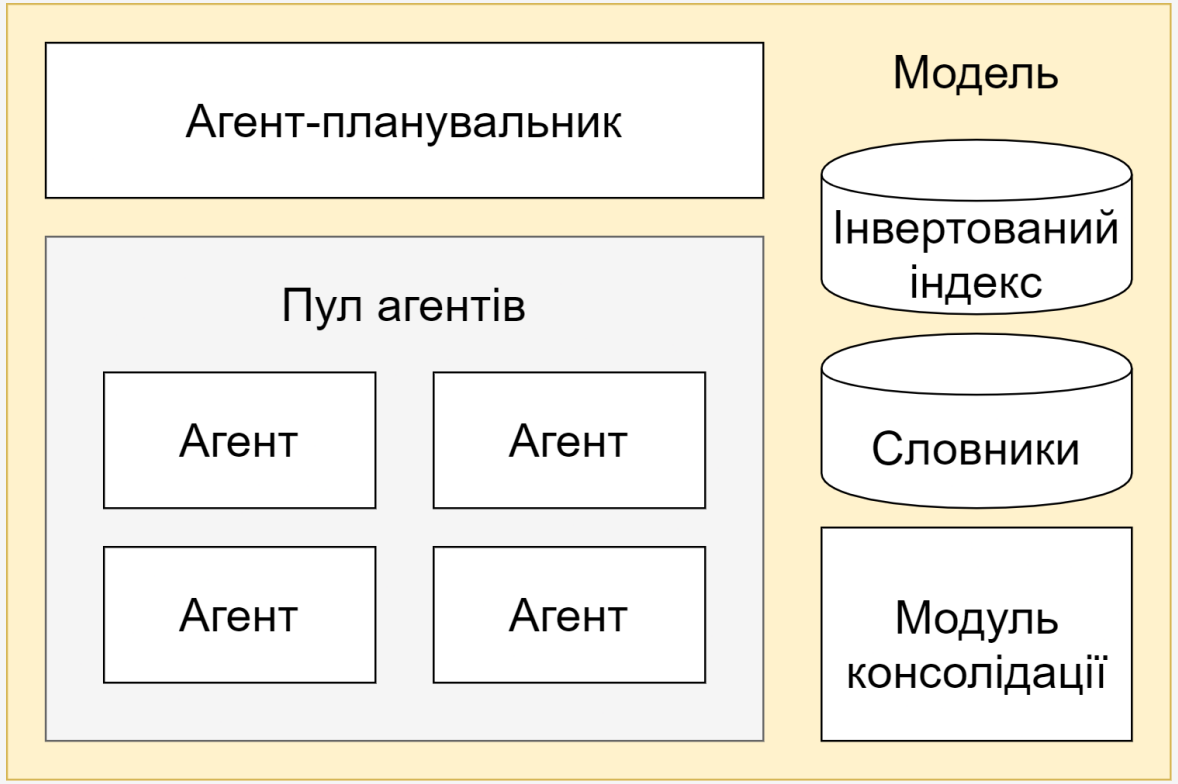
- **TextRank** – алгоритм, котрий базується на підході PageRank та представляє собою алгоритм витягу ключових фраз на основі обходу графу

Обидва алгоритми оцінюють важливість слова для одного документу

Результати RAKE та TextRank

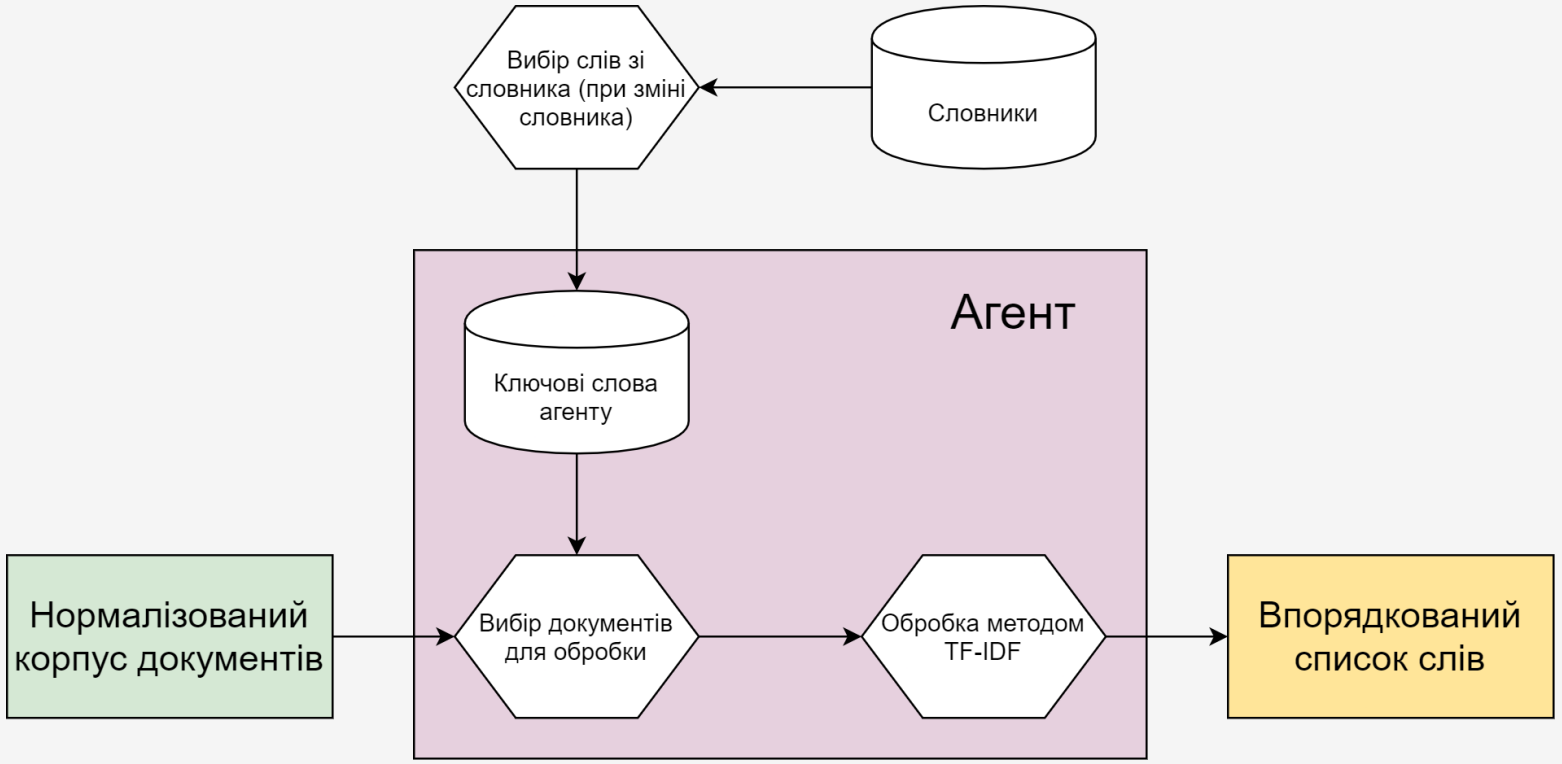
RAKE		TextRank	
Словосполучення	**Відносна Оцінка**	**Словосполучення**	**Відносна Оцінка**
Semiconductor Diodes begin conducting electricity	20.409	**Electronic Device**	0.119
Impatt diodes exhibit negative resistance	19.909	**Power Device**	0.110
Perform many different functions	16.000	**Such Device**	0.096
Reverse direction suddenly drops	14.000	**Electronic System**	0.090
Impatt diodes	9.909	**Crystalline Solids**	0.088
Doping impurities introduced	9.000	**Wide Application**	0.081
Zener Diodes	8.909	**Communications**	0.060
Varactor Diodes	8.909	**Integration**	0.058
Avalanche Diodes	8.833	**Computing**	0.055
Reverse Voltage Across	8.500	**Transistors**	0.055

Архітектура розробленої моделі



- Мною була розроблена архітектуру моделі, котра складається з компонентів, перелічених на рисунку.
- Аналізом текстів займаються агенти, котрі використовують метод TF-IDF.

Робота агенту

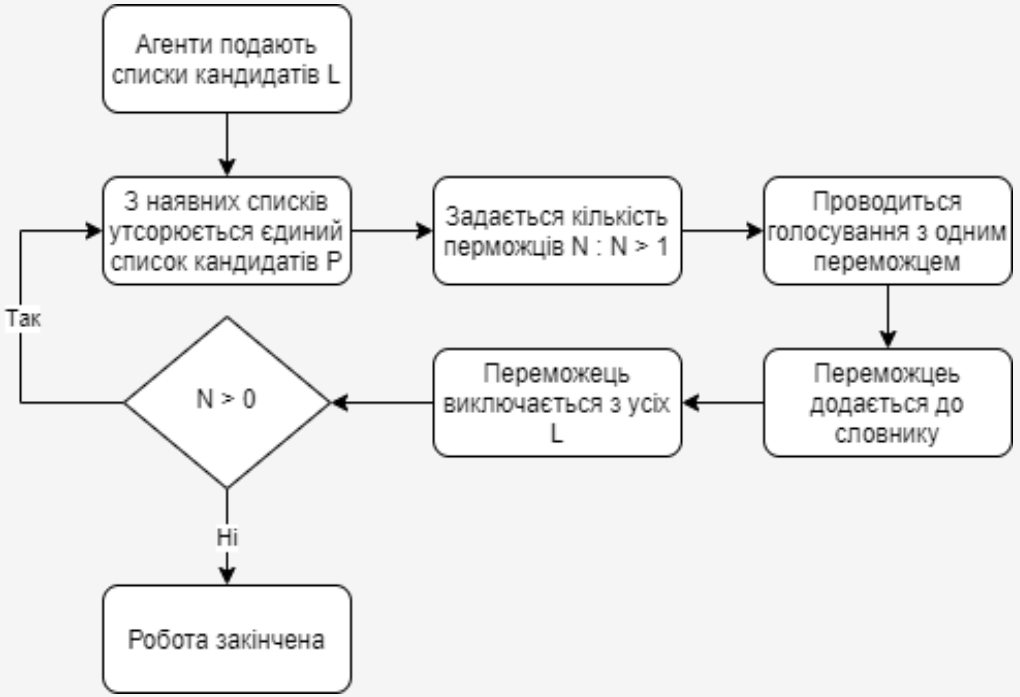


- Після отримання вхідних даних від планувальника, агент починає обробку даних.

Голосування агентів

- Для голосування агентів було обрано метод Шульце з одним переможцем та модифіковано його використання.

Метод	Кандидати		Час	Результати
	Всього	Переможців		
Модифікований метод Шульце	22	2	0.12 с	**gravity, interaction, quantum, star**
Метод Шульце з декількома переможцями	22	2	1382.1 с	**gravity, interaction, quantum, star**



Результати роботи системи

Робота системи перевірялася на двох наявних наборах даних з областей:

- Фізика
- Біологія

Предметна область – фізика, 11 документів, 4185 слів

```
INFO:root:Agents have voted for ['gravity', 'interaction', 'quantum', 'star'].
```

Предметна область – біологія, 6 документів, 6182 слів

```
INFO:root:Agents have voted for ['brain', 'vertebrate', 'structure', 'size'].
```

Засоби розробки

Для проведення аналізу методів та розробки було використано:

План подальшої роботи:

- Дослідження залежності ефективності роботи системи від розміру словника та кількості агентів в системі. Дослідження коефіцієнту Амдала для розробленої архітектури.
- Дослідження роботи системи на великих наборах текстових даних та тестування навантаження MAC.
- Дослідження впливу розміру корпусу одного агенту на точність роботи.

Висновки:

За результатами роботи можемо надати такі висновки:

- Частотний аналіз з використанням правильного методу нормалізації є найбільш ефективним методом пошуку ключових слів на основі словника.
- Векторизація не є надійним способом знаходження подібних слів в корпусі даних зі слабким контекстом.
- Правильний розподіл ключових слів між агентами в MAC досить сильно впливає на результат роботи системи.

ДЯКУЮ ЗА УВАГУ!

Студент групи ДА-62
Сиротюк Олександр Васильович