Для порівняння речень на подібність в даній реалізації використовувались Jaccard similarity та Cosine similarity. Дані реалізації можуть також використовуватись для порівння коротких текстових повідомлень і текстів загалом.

Для обчислення Jaccard similarity визначається як розмір перетину, поділений на величину об'єднання двох множині і тому для її підрахунку необхідно (кроки 1-4 виконуються для кожного речення окремо):

1. Токенізувати обидва речення за словами
2. Видалити стоп слова
3. Видалити знаки пунктуації
4. Застосувати стемінг або лематизацію до кожного слова речень
5. Знайти симетричну різницю множин слів речень
6. Поділити знайдену симетричну різницю на розмір об'єднання множин слів речень без повторів.

Основою підходу Cosine similarity є розгляд векторів, які побудовані на основі речень, і визначення їх зміщення відносно один одного.

Для обчислення Cosine similarity необхідно:

1. Токенізувати речення за словами
2. Видалити стоп слова
3. Видалити знаки пунктуації
4. Обчислити об'єднання без повторів множин інформативних слів речень
5. Сформувати характеристичні вектори для речень:
   шляхом проходу по отриманому об'єднанню і перевірці входження слова в речення. Якщо слово з об'єднання на і-тому кроці входить в речення, то на і-тій позиції вектора ставиться 1, а інакше 0.
6. Обчислити Cosine similarity за наступною формулою

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

В рамках роботи було досліджено наступні випадки:

- використання синонімів під час обчислення Cosine similarity шляхом перевірки на 5 кроці під час формування векторів входження не просто слова, а і його синонімів. Використання синонімів призвело до покращення результатів. Так для двох схожих речень (в подальших випадках дані речення є вхідними даними для наведених результатів реалізації)

```
sentence1 ="I love funny books"
sentence2 ="Calm is a funny book, which I like"
```

обчислення з використанням синонімів дали більш реалістичний прогноз (нижче наведено результат роботи програми).

```
cosine similarity with using synonims:  0.6123724356957946
cosine similarity without using synonims:  0.4082482904638631
Jaccard similarity:  0.2727272727272727
```

- Порівняно Jaccard similarity і Cosine similarity

```
cosine similarity with using synonims:  0.6123724356957946
cosine similarity without using synonims:  0.4082482904638631
Jaccard similarity:  0.2727272727272727
```

як видно з результату роботи програми Cosine similarity дає більш правдоподібні результати навіть без використання покращення з синонімами ніж Jaccard similarity

- спроба модифікації обчислення Cosine similarity шляхом використання 4 кроку (1-3 кроки збігаються) Jaccard similarity алгоритму під час попередньої обробки речень і вже на основі отриманих множин слів після стемінгу та/або лемітизації будувати характеристичні вектори і досліджувати їх на подібність і також досліджувати міру їх відхилення один від одного. Це дало наступні результати

```
cosine similarity with using synonims:  0.5
cosine similarity without using synonims:  0.6123724356957946
```

тобто покращення Cosine similarity без використання синонімів (в минулому прикладі можна побачити результат виконання без використання стемінгу та/або лемітизації, де Cosine similarity без використання синонімів складала 0,4082482904638631). Це відбулося за рахунок зменшення розмірності векторів, а отже міри їх відмінності один від одного.