Article Summarizer

Лиляна Видева, ИИОЗ, 26116

* Хранилище

<https://github.com/LiliVideva/NLP-IR-Project-2020>

* Обща характеристика

Темата на проекта се концентрира около обработката на информация в набор от статии от Wikipedia за американските президенти. Частта, която е в областта на обработката на естествен език, наречена Article Summarizer, цели да предостави възможност за резюмиране на съдържанието на документите с помощта на TextRank.

* Описание на работата на програмата

Опоменатите стъпки се изпълняват върху всяка една статия:

1. Превръщане на всяка в статия в един низ
2. Определяне на отделните изречения в низа
3. Изчисляване на резултат за всяко изречение
   1. Назначаване на тегла на думите
   2. Класиране на изреченията с помощта на TextRank
   3. Откриване на най-подходящо обобщаващите изречения
   4. Дефиниране на изреченията с най-висок резултат
4. Сглобяване на резюме от „най-добрите“ изречения

* Описание на алгоритъма TextRank

Изпълнението на подхода се концентрира върху прилагането на следната формула:

където:

* е претегленото сходство(weighted similarity) на върха
* е множеството от предшествениците на
* е множеството от наследниците на
* е теглото от до
* е амортизиращ фактор, който играе има за цел да симулира вероятността, с която се прескача от един връх на произволен друг връх, и обикновено приема стойност

Имплементацията следва следния ред от действия:

1. Изчисляване на сходството между съседни изречения на базата на общи думи и тяхната дължина и добавяне на взаимовръзки между тях (Проверка дали не са „стоп“ думи)
2. Изчисляване на сумата по множеството от предшествениците
3. Определяне на крайните резултати по формулата спрямо предварително зададен брой итерации и максимална грешка и сортирането им в нарастващ ред