Лабораторная работа

«Разработка системы автоматического реферирования документов»

Цель работы – освоить на практике основные принципы автоматического реферирования документов.

Весовые коэффициенты значимости терминов

Функция автоматического реферирования – дать сжатое представление текстовой информации, позволяющее пользователю экономить время при поиске и отборе необходимой информации, т.е. «отсеивать» менее значимую информацию.

Можно выделить следующие типы рефератов:

- реферат в виде списка ключевых слов
- классический реферат
- структурированный реферат
- запросно-ориентированный реферат

Реферат в виде ключевых слов представляет собой список, возможно иерархический (в виде дерева) наиболее информативных слов и словосочетаний (именных групп) обрабатываемого документа. Такой реферат позволяет пользователю понять основные темы описанные в документе. Например: *лазер, лазерный луч, синий лазер, красный лазер, устройство* и т.д. Или в иерархическом виде*:*

лазер

лазерный луч

синий лазер

красный лазер

устройство

и т.д.

Классический реферат – это набор наиболее информативных предложений текста, возможно трансформированных (удаление вводных конструкций, замена анафоричных местоимений и т.д. с целью улучшения связности реферата и уменьшения его объема).

При построении классического реферата методом sentence extraction необходимо:

1. Вычислить веса слов документа.

При этом слова из латинских букв, числа, стоп-слова - не учитываются. Базовый вес слова вычисляется по формуле TF*IDF.

2. Вычислить веса предложений согласно формулам, приведенным ниже.
3. Осуществить генерацию реферата.

Этап генерации представляет собой выбор из исходного текста определенного количества предложений с наибольшим весом в той последовательности, в которой они идут в тексте. Рекомендуемый размер реферата 10 предложений.

Вес каждого предложения S_i вычисляется произведением значений функций приведенных ниже.

Функции, характеризующие положение предложения в документе $\mathrm{Posd}(S_i)$ и положение в абзаце $\mathrm{Posp}(S_i)$:

$$\text{Posd}(S_i) = 1 - \frac{BD(S_i)}{|D|}$$
$$\text{Posp}(S_i) = 1 - \frac{BP(S_i)}{|P|}.$$

где

|D|- число символов в документе D, содержащем предложение S_i;
$BD(S_i)$ – количество символов до S_i в $D(S_i)$;
|P| - количество символов в абзаце P, содержащем предложение S_i;
$BP(S_i)$ – количество символов до S_i в абзаце.

Модифицированная TFIDF функция:

$$\text{Score}(S_i) = \sum_{t \in S_i} tf(t, S_i) \cdot w(t, D).$$

$tf(t, S_i)$- частота термина t в предложении S_i;

$$w(t, D) = 0.5 \left(1 + \frac{tf(t, D)}{tf_{max}(D)}\right) \cdot \log\left(\frac{|DB|}{df(t)}\right).$$

$tf(t, D)$- частота термина t в документе D;
$df(t)$- количество документов, с термином t;
$tf_{max}(D)$ - максимальная частота термина в документе D;
|DB| - количество документов.

Требования к разрабатываемой системе

- ✓ на входе – на входе – текстовые документы одинакового размера (например, 10 страниц формата А4), содержащие тексты из предметных областей на естественных языках согласно варианту подлежащие процедуре автоматического реферирования;
- ✓ на выходе – активная ссылка на исходный документ и построенный, в соответствии с вариантом реферат документа, состоящий из 2-х разделов:
  - 1 - классического реферата и реферата в виде списка ключевых слов (по методу Sentence extraction);
  -2 – реферата, построенного с применением машинного обучения (ML).
- ✓ наличие средств сохранения в файл и распечатки полученной на выходе информации;
- ✓ интерфейс системы должен быть предельно простым и доступным для пользователей любого уровня, содержать понятный набор инструментов и средств, а также help-средства.

№ варианта	Язык текста	Методика	Предметная область
1	Русский, Английский	Sentence extraction+ ML	Научные статьи по computer science, Сочинения по литературе
2	Французский, Немецкий	Sentence extraction+ ML	Научные статьи по computer science, Сочинения по литературе
3	Испанский, Итальянский	Sentence extraction+ ML	Научные статьи по computer science, Сочинения по литературе
4	Русский, Немецкий	Sentence extraction+ ML	Научные статьи по computer science, Сочинения по литературе
5	Русский, Итальянский	Sentence extraction+ ML	Научные статьи по computer science, Сочинения по литературе
6	Французский, Итальянский	Sentence extraction+ ML	Научные статьи по computer science, Сочинения по литературе
7	Французский, Английский	Sentence extraction+ ML	Научные статьи по computer science, Сочинения по литературе
8	Испанский, Немецкий	Sentence extraction+ ML	Научные статьи по computer science, Сочинения по литературе
9	Испанский, Английский	Sentence extraction+ ML	Научные статьи по computer science, Сочинения по литературе
10	Русский, Английский	Sentence extraction+ ML	Научные статьи по медицине, Критика предметов изобразительного искусства
11	Французский, Немецкий	Sentence extraction+ ML	Научные статьи по медицине, Критика предметов изобразительного искусства
12	Испанский, Итальянский	Sentence extraction+ ML	Научные статьи по медицине, Критика предметов изобразительного искусства
13	Русский, Немецкий	Sentence extraction+ ML	Научные статьи по медицине, Критика предметов изобразительного искусства
14	Русский, Итальянский	Sentence extraction+ ML	Научные статьи по медицине, Критика предметов изобразительного искусства
15	Французский, Итальянский	Sentence extraction+ ML	Научные статьи по медицине, Критика предметов изобразительного искусства
16	Французский, Английский	Sentence extraction+ ML	Научные статьи по медицине, Критика предметов изобразительного искусства
17	Испанский, Немецкий	Sentence extraction+ ML	Научные статьи по медицине, Критика предметов изобразительного искусства
18	Испанский, Английский	Sentence extraction+ ML	Научные статьи по медицине, Критика предметов изобразительного искусства
19	Русский, Английский	Sentence extraction+ ML	Научные статьи по медицине, Критика предметов

			изобразительного искусства
20	Французский, Немецкий	Sentence extraction+ ML	Научные статьи по медицине, Критика предметов изобразительного искусства
21	Испанский, Итальянский	Sentence extraction+ ML	Научные статьи по медицине, Критика предметов изобразительного искусства
22	Русский, Немецкий	Sentence extraction+ ML	Научные статьи по медицине, Критика предметов изобразительного искусства
23	Русский, Итальянский	Sentence extraction+ ML	Научные статьи по медицине, Критика предметов изобразительного искусства
24	Французский, Итальянский	Sentence extraction+ ML	Научные статьи по медицине, Критика предметов изобразительного искусства
25	Французский, Английский	Sentence extraction+ ML	Научные статьи по медицине, Критика предметов изобразительного искусства
26	Испанский, Немецкий	Sentence extraction+ ML	Научные статьи по медицине, Критика предметов изобразительного искусства
27	Испанский, Английский	Sentence extraction+ ML	Научные статьи по медицине, Критика предметов изобразительного искусства

В отчет по работе необходимо включить:

- ✓ информацию о тестовой коллекции документов (может быть сформирована на основании многоязычных текстов из Wikipedia, Medline, других тематических ресурсов);
- ✓ описание структуры разработанной системы;
- ✓ описание структур данных, использованных для хранения входной и выходной информации;
- ✓ описание алгоритма построения реферата (в текстовом и графическом виде);
- ✓ результаты тестирования системы;
- ✓ оценку полученных результатов (оценка точности и затраченного времени т.д. для текстов разных предметных областей и естественных языков);
- ✓ описание и особенности применения готовых к использованию компонент (библиотек, классов, фреймворков и т.п.) в случае их использования в работе;
- ✓ сравнить результаты, полученные различными методами;
- ✓ выводы по работе и перспективам применения разработанной системы.