Поиск неточных повторов

в программной документации

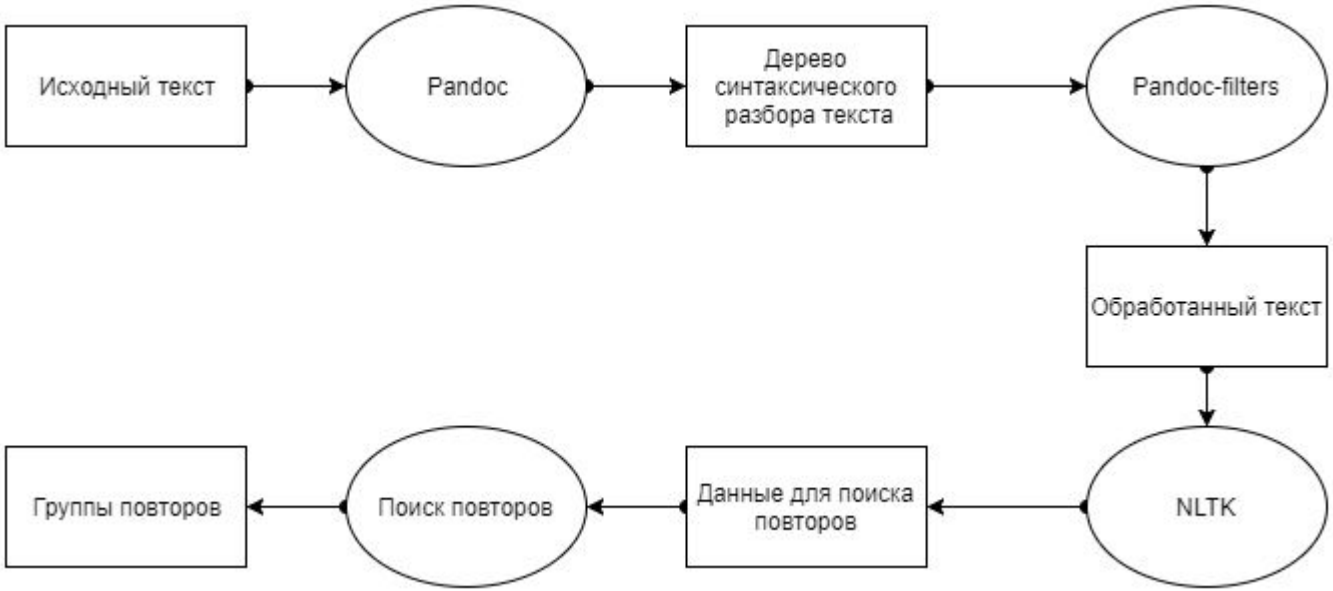
Кантеев Леонид Дмитриевич, 371

Научный руководитель: д.т.н., доцент Кознов Д. В

Постановка задачи

- Улучшить существующий новый алгоритм нечёткого поиска на основе N-грамм
- Интегрировать алгоритм в Docline/Duplicate Finder.

Алгоритм


Исходный текст
Pandoc
Дерево синтаксического разбора текста
Pandoc-filters
Обработанный текст
NLTK
Данные для поиска повторов
Поиск повторов
Группы повторов

Поиск повторов

```
1.  fragments ← list of document sentences
2.  groups ← ∅
3.  for f1 ∈ fragments do
4.    if Involved(f1) then
5.      continue
6.    curFr ← f1
7.    curGr ← ∅
8.    for f2 ∈ {f ∈ fragments|¬Before(f, f1)} do
9.      if Suitable(curGr, f2) then
10.       Append(curGr, f2)
11.   while #curGr > 1 do
12.     expandedGr ← ∅
13.     for f3 ∈ curGr do
14.       newFr ← Expand(f3)
15.       if Suitable(expandedGr, newFr) then
16.         Append(expandedGr, newFr)
17.         continue
18.     if expandedGr < 2 then
19.       break
20.     else
21.       curGr ← expandedGr
22.   Append(grps, curGr)
23. for ∀gr ∈ grps, if gr = 1 remove gr from grps
24. return grps
```

Результаты

- Алгоритм расширен на мультипоиск
- Улучшено время работы
- Улучшено качество повторов
- Сделан парсинг разметки
- Алгоритм интегрирован в Docline/Duplicate Finder

Качество повторов

	Процент переиспользования	Нерелевантных повторов	Релевантных повторов	Процент осмысленного переиспользования
Было	19,00%	24,50%	61,00%	10,80%
Стало	19,30%	21,68%	78,32%	13,51%

Производительность

Было:

Документ	Размер, кб	Время работы, с	Средняя скорость работы, кб/с
LKD	892	318	2.80
Zend	2924	1328	2.20
DocBook	686	121	5.66
SVN	1810	1028	1.76
CProj	164	10	16.40

Стало:

Документ	Размер, кб	Время работы, с	Средняя скорость работы, кб/с
LKD	892	20.3	43.94
Zend	2924	56.2	52.02
DocBook	686	8.1	84.69
SVN	1810	91.1	19.86
CProj	164	2.1	78.09