КЛАССИФИКАЦИЯ СКРЕП

N00. Разрывные скрепы

N01. ИДЕНТИФИКАЦИЯ

N02. СРАВНЕНИЕ

N03. ПРИЧИНА

N04. ВОЗМОЖНОСТЬ

N05. ПРОТИВОПОСТАВЛЕНИЕ, ПРОТИВИТЕЛЬНОСТЬ, УСТУПИТЕЛЬНОСТЬ

N06. ОБОБЩЕНИЕ-КОНКРЕТИЗАЦИЯ

N07. ОБОСОБЛЕНИЕ-ОГРАНИЧЕНИЕ

N08. ЦЕЛЬ

N09. ВРЕМЯ

N10а. ПРЕДСКАЗУЕМОСТЬ, СЧЁТНОСТЬ, ПОСТОЯНСТВО, УСТОЙЧИВОСТЬ

N10б. НЕПРЕДСКАЗУЕМОСТЬ

N11. УСЛОВИЕ

N12. ВВОД ТЕМЫ (ИСТОЧНИК ИНФОРМАЦИИ)

N13. МЕСТО

N14. ОБЪЕДИНЕНИЕ

ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

$C_j - коннектор\_номер\_j$

$T_k - текст\_номер\_k$

$L(T_k) - длина\_текста\_T_k\_(количество\_словоформ)$

$n_j(T_k) - количество\_появлений\_коннектора\_C_j\_в\_тексте\_T_k$

$частота\_коннектора\_C_j\_в\_тексте\_T_k$:

$$f_j(T_k) = \frac{n_j(T_k) * 1000000}{L(_k)} =$$

$$\frac{количество\_появлений\_коннектора\_C_j\_в\_тексте\_T_k, \_умноженное\_на\_1000000}{длина\_текста\_T_k}$$

$средняя\_частота\_коннектора\_C_j\_в\_текстах\_некоторого\_автора =$

$$f_j^* == \frac{суммарное\_количество\_появлений\_коннектора\_C_j\_в\_текстах\_автора}{суммарная\_длина\_текстов\_автора} * 1000000 =$$

$$\frac{n_j(T_1) + n_j(T_2) + ... + n_j(T_n)}{L(T_1) + L(T_2) + ... + L(T_n)} * 1000000 = \frac{\sum\limits_k n_j(T_k)}{\sum\limits_k L(T_{k)}} * 1000000 = \sum\limits_k f_j(T_k) * \frac{L_k}{\sum\limits_m L(T_m)}$$

Для каждого текста вычисляется последовательность частот:

$f_1(T_k)$, $f_2(T_k)$, $f_3(T_k)$, **...**

Мы можем рассматривать эти числа как координаты точки T_k в многомерном пространстве.

Текстам каждого автора соответствует некоторое множество – облако точек. Удобно считать, что в каждой точке сосредоточена масса, равная объёму текста. Тогда числа f_j^* – это координаты центра тяжести облака точек. Мы можем для наглядности считать, что точки каждого автора закрашены своим цветом.

Суммы $\sum_j f_j(T)$ для разных текстов одного и того же автора могут изменяться в довольно широком интервале. При этом интервалы, соответствующие текстам разных авторов, могут сильно перекрываться.

РАССТОЯНИЕ МЕЖДУ ТЕКСТАМИ

Расстояние между текстами T_a и T_b=

$$\sum_j | f_j(T_a) - f_j(T_b) |$$

Расстояние между текстом и множеством отличных от него текстов некоторого автора

$$\sum_j | f_j(T) - f_j^* |$$

Расстояние от текста до облака точек, соответствующих текстам некоторого автора, определяется как расстояние от точки, соответствующей тексту, до центра тяжести облака точек. При этом точка, соответствующая исследуемому тексту, исключается из облака точек, соответствующих текстам автора, если она ему принадлежала.

Таблица авторов

ТАБЛИЦА ПОДГОТОВЛЕННЫХ ДЛЯ ОБРАБОТКИ ТЕКСТОВ

Номер текста	Номер предложения	Номер слова	Слово	Знаки препинания после слова
92	1	1	Повести	,
92	1	2	Изданные	
92	1	3	пасичником	
92	1	4	Рудым	
92	1	5	Паньком	
92	2	1	Часть	
92	2	2	ПЕРВАЯ	
92	3	1	Предисловие	«
92	4	1	Это	
92	4	2	что	
92	4	3	за	
92	4	4	невидаль	: «
92	4	5	Вечера	
92	4	6	на	
92	4	7	хуторе	
92	4	8	близ	
92	4	9	Диканьки	»?

ПРИМЕР ПЕРЕСЕЧЕНИЯ КОННЕКТОРОВ

933	Нет$_{1}$, матушка$_{2}$ не$_{3}$ обижу$_{4}$, – говорил$_{5}$ он$_{6}$, **а$_{7}$ между$_{8}$ тем$_{9}$** отирал$_{10}$ рукою$_{11}$ пот$_{12}$, который$_{13}$ в$_{14}$ три$_{15}$ ручья$_{16}$ катился$_{17}$ по$_{18}$ лицу$_{19}$ его$_{20}$.	7; 8; 9;	1	ТОТ (2-15), КОТОРЫЙ	9; 13;	0
				МЕЖДУ ТЕМ	8; 9;	2
				НЕ (2-15), А	3; 7;	1