ЛАБОРАТОРНАЯ РАБОТА №2

Дата сдачи - 19 октября 2022

На базе собранного корпуса и созданого частотного словаря (ЧС содержит только слова, никаких цифр, спецсимволов и слов, начинающихся со спецсимволов) провести статистические исследования в соответствии с материалами лекции 3 (файл Лекция 3.pdf в папке "Спецкурс/Лекции"):

Задание 1. Закон Ципфа

– рассчитать показатели с построением графика,

– проверить выполнение формулы (определить коэффициент Ципфа) и сделать вывод о выполнении закона

$$f \cdot r = c$$

где ***f*** – частота встречаемости слова в тексте;

r – ранг слова в списке;

c – эмпирическая постоянная величина (коэффициент Ципфа).

Задание 2. 2-й закон Ципфа

– рассчитать константу нормирования ***B*** в соответствии с формулой для первых 10 максимальных значений частоты:

$$N(f) = \frac{B}{f^b},$$

где ***N(f)*** – количество различных слов, каждое из которых используется в тексте ***f*** раз,

B – константа нормирования;

b — количество слов.

– построить график, где X - частота слова, Y – число слов данной частоты, проверить, будет форма кривой будет неизменной.

Задание 3. Эмпирический закон Ципфа

– на основе списка служебных слов из созданного частотного словаря проверить выполнение эмпирического закона Ципфа*: длина слова обратно пропорциональна его частоте.*

Задание 4. Расчет коэффициента *D* Жуйана

Рассчитать коэффициент для полученного частотного словаря и результат оформить в виде таблицы (слово, относительная частота, ранг, D).

Коэффициент ***D*** отражает равномерность распределения частот в разных сегментах корпуса и вычисляется по следующей формуле:

$$D = 100 \times (1 - \frac{\sigma}{\mu\sqrt{n-1}})$$

где μ – средняя частота слова по всему корпусу,

σ – среднее квадратичное отклонение этой частоты на отдельных документах,

n – число документов, в которых встречается это слово.

Для подсчета коэффициента Жуйана корпус разбивается на n равных сегментов (например, на 4 части, размером приблизительно в 250 тыс. слов каждый).

Пример

Пусть исследуемые тексты разбиты на 4 сегмента, каждый размером по 1 миллиону слов. Некоторое слово, например, «коэффициент» встречается в этих сегментах соответственно 10, 11, 8 и 3 раза. Тогда $\mu_1 = 10$, $\mu_2 = 11$, $\mu_3 = 9$, $\mu_4 = 3$ употребления на миллион. Среднее значение

$$\mu = \frac{10+11+8+3}{4} = 8.$$

Среднеквадратическое отклонение

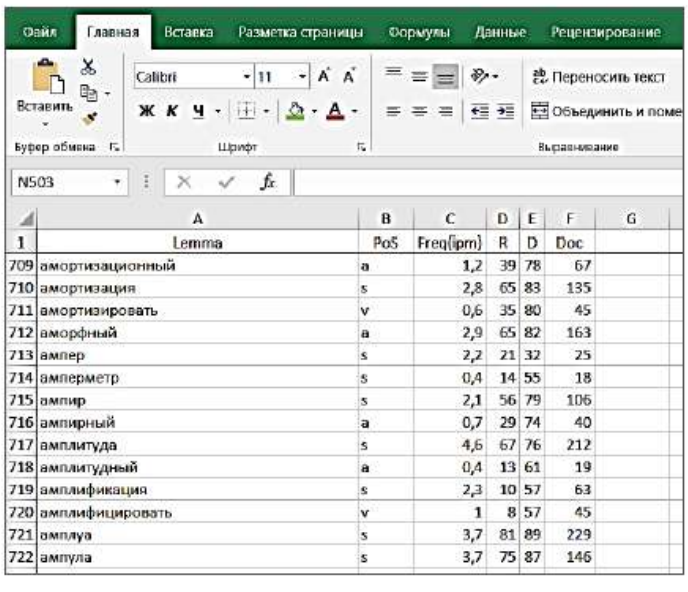
$$\sigma = \sqrt{\frac{(10-8)^2+(11-8)^2+(8-8)^2+(3-8)^2}{4}} = \sqrt{\frac{4+9+0+25}{4}} = \sqrt{9,5} = 3,08.$$

Тогда Коэффициент Жуайна

$$D = 100\left(1 - \frac{3,08}{8\sqrt{4-1}}\right) = 100(1-0,22) = 78.$$

Пример таблицы:

Коэффициент Жуйана D



	A	B	C	D	E	F	G
1	Lemma	PoS	Freq(ipm)	R	D	Doc	
709	амортизационный	a	1,2	39	78	67	
710	амортизация	s	2,8	65	83	135	
711	амортизировать	v	0,6	35	80	45	
712	аморфный	a	2,9	65	82	163	
713	ампер	s	2,2	21	32	25	
714	амперметр	s	0,4	14	55	18	
715	ампир	s	2,1	56	79	106	
716	ампирный	a	0,7	29	74	40	
717	амплитуда	s	4,6	67	76	212	
718	амплитудный	a	0,4	13	61	19	
719	амплификация	s	2,3	10	57	63	
720	амплифицировать	v	1	8	57	45	
721	ампула	s	3,7	81	89	229	
722	ампула	s	3,7	75	87	146	

ТРЕБОВАНИЯ К ОФОМЛЕНИЮ

Отчет по лабораторной работе оформляется в виде doc-файла, удовлетворяющего требованиям, описанным ниже. Отчет предоставляется в электронном виде (загружается на EDU).

1. Структура отчета

Отчет должен иметь следующую структуру:

1) титульный лист;

2) содержание;

3) описание заданий – постановка и решение.

2. Правила оформления

Набор текста осуществляется с использованием текстового редактора Word. При этом рекомендуется использовать шрифты типа Times New Roman размером 14 пунктов. Межстрочный интервал должен составлять 18 пунктов.

Устанавливаются следующие размеры полей: верхнего и нижнего - 20 мм, левого - 30 мм, правого - 10 мм.

Заголовки разделов и названия заданий печатают прописными буквами полужирным шрифтом с размером на 1-2 пункта больше, чем в основном тексте, располагая их по центру. В конце заголовков точку не ставят. Названия заданий должны соответствовать указанным в лабораторной работе.

Расстояние между заголовком и текстом должно составлять 2-3 межстрочных интервала.

Каждую структурную часть и каждое задание следует начинать с нового листа.

Графики оформляются как рисунки. Иллюстрации и таблицы обозначают соответственно словами "рисунок" и "таблица" и нумеруют последовательно в пределах всего текста. На все таблицы и иллюстрации должны быть ссылки в тексте.

Иллюстрации имеют подпись, располагаемую по центру страницы под рисунком. Подпись содержит слово "Рисунок", номер и наименование, отделенное тире от номера. Точку в конце нумерации и наименований иллюстраций не ставят. Не допускается перенос слов в наименовании рисунка. Слово "Рисунок", его номер и наименование иллюстрации печатают полужирным шрифтом с уменьшенным на 1-2 пункта размером шрифта.

Таблицу следует располагать непосредственно после текста, где она упоминается впервые, или на следующей страниц, если не помещается после текста.

Каждая таблица должна иметь краткий заголовок, который состоит из слова "Таблица", ее порядкового номера и названия, отделенного от номера знаком тире. Заголовок следует помещать над таблицей слева, без абзацного отступа.

Пример оформления титульного листа

Белорусский государственный университет
Факультет прикладной математики и информатики

Кафедра информационных систем управления

Лабораторная работа №___

Выполнил
студент № курса № группы
ФИО

Минск, 2022