Объединение данных из разных таблиц с помощью индексов

Предположим, мы хотим объединить две имеющиеся у нас таблицы в одну. Как это сделать? Сегодня мы разберем это на примере.

Например, мы имеем частотные списки слов в из двух корпусов, КРУТ и Русского учебного корпуса - частотные списки обрезаны по частоте 3. (скачать таблицы)

Таблица corst_freq

	A	B	C	D
1	Слово	Частота	IPM	
2	в	35889	11520,56	
3	и	31435	10090,81	
4	на	13212	4241,124	
5	не	11811	3791,395	
6	с	11363	3647,585	
7	что	10265	3295,121	
8	как	7396	2374,156	
9	к	6790	2179,627	
10	а	5924	1901,636	

Таблица rlc_freq

	A	B	C	D
1	Слово	Частота	IPM	
2	в	35889	11520,56	
3	и	31435	10090,81	
4	на	13212	4241,124	
5	не	11811	3791,395	
6	с	11363	3647,585	
7	что	10265	3295,121	
8	как	7396	2374,156	
9	к	6790	2179,627	
10	а	5924	1901,636	

Мы хотим получить общую таблицу, где данные из двух таблиц сведены вместе и их можно сравнить.

	A	B	C
1	Слово	IPM_corst	IPM_rlc
2	в	11520,56	37993,23
3	и	10090,81	38334,70
4	на	4241,12	13265,41
5	не	3791,40	12860,17
6	с	3647,58	9178,43
7	что	3295,12	16305,40
8	как	2374,16	8273,61
9	к	2179,63	4742,01
10	а	1901,64	3449,22
11	для	1814,32	5009,07

Что нужно делать:

Шаг 1.

Добавим в таблицу corst_freq столбец "Номер строки, на которой слово стоит в rlc_freq".

	A	B	C	D	E	F	G	H
1	Слово	Частота	IPM	Номер строки, на которой слово стоит в rlc_freq				
2	в	35889	11520,56		(Ctrl)			
3	и	31435	10090,81					
4	на	13212	4241,124					
5	не	11811	3791,395					
6	с	11363	3647,585					
7	что	10265	3295,121					
8	как	7396	2374,156					
9	к	6790	2179,627					
10	а	5924	1901,636					

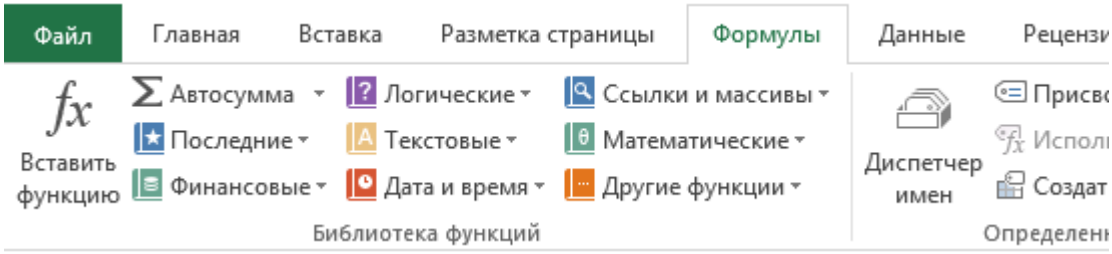
Шаг 2.

Затем для каждого слова в таблице corst_freq нужно:

- найти его в столбце "Слово" таблицы rlc_freq

- определить номер строки, на которой он стоит
- записать этот номер в столбец "Номер..." таблицы corst_freq

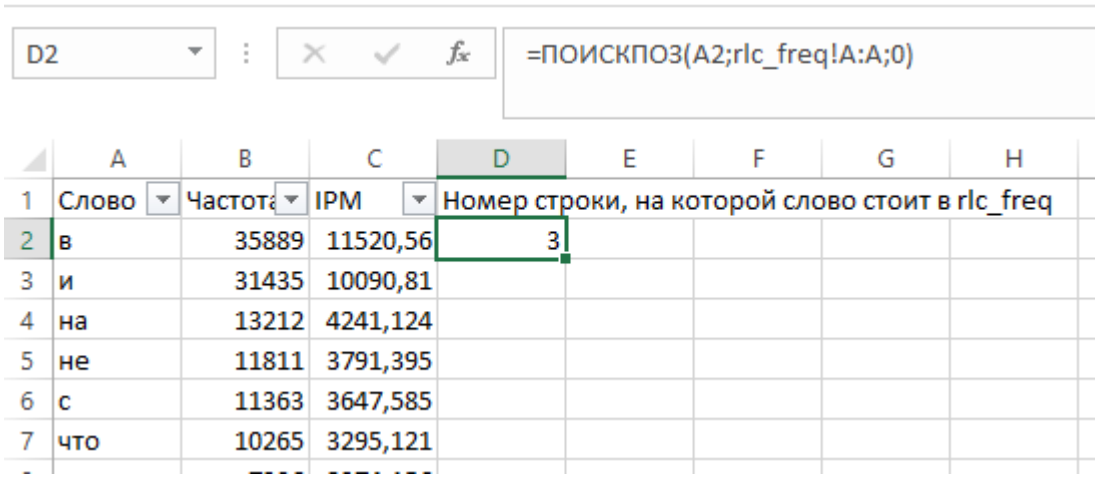
Чтобы в Excel выполнить эти три пункта, на вкладке Формулы найдем раздел "Ссылки и массивы" и в нем формулу "ПОИСКПОЗ" (в английской версии "MATCH").



В открывшемся окне вставки формул укажем:

- Искомое_значение: кликните на ячейку с лексемой слева.
- Просматриваемый массив: затем перейдите на лист таблицы rlc_freq и выделите столбец "Слово"
- Тип сопоставления: 0 (обозначает точное совпадение).

Нажмите OK.



В ячейке должен отобразиться номер строки, на которой искомое слово стоит в таблице rlc_freq. Проверьте (с помощью поиска), что номер правильный.

Скопируем ячейку с формулой и вставим ее в том же столбце напротив всех остальных слов (можно дважды кликнуть на правый нижний угол заполненной ячейки, чтобы ее формула автоматически растянулась на весь столбец).

	A	B	C	D
1	Слово	Частота	IPM	Номер строки
2	в	35889	11520,56	3
3	и	31435	10090,81	2
4	на	13212	4241,124	5
5	не	11811	3791,395	6
6	с	11363	3647,585	8
7	что	10265	3295,121	4
8	как	7396	2374,156	9
9	к	6790	2179,627	15
10	а	5924	1901,636	20
11	для	5652	1814,323	14
12	по	5634	1808,545	16
13	о	4912	1576,779	13
14	из	4365	1401.189	17

NB Пересчет значений ячеек может занять некоторое время, особенно для больших таблиц. Если после пересчета значений в каких-то ячейках появится #Н/Д (в английской версии - #N/A), это означает, что строка с таким словом не найдена.

Шаг 3

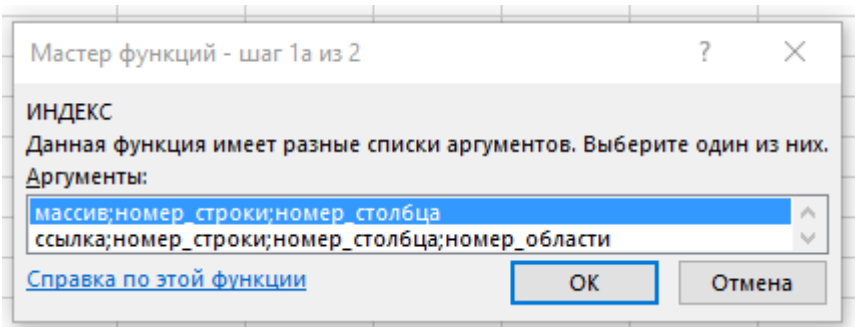
Добавим в таблицу corst_freq столбец "IPM_rlc".

	A	B	C	D	E
1	Слово	Частота	IPM	Номер стр	IPM_rlc
2	в	35889	11520,56	3	
3	и	31435	10090,81	2	
4	на	13212	4241,124	5	
5	не	11811	3791,395	6	

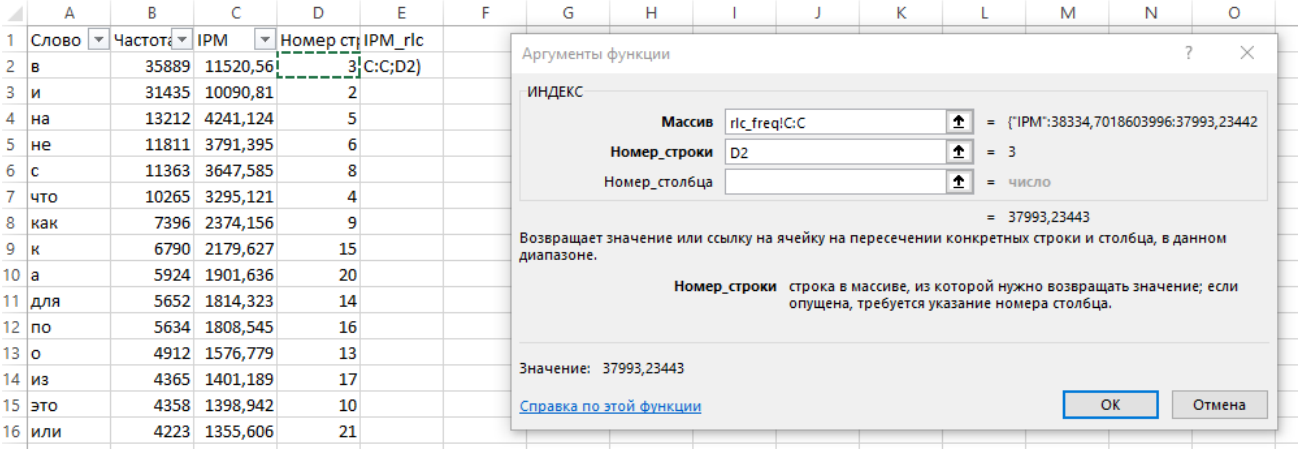
Шаг 4

На вкладке Формулы в разделе "Ссылки и массивы" найдем формулу "ИНДЕКС" (в английской версии "INDEX"), и в открывшемся мастере формулы:

- выберите первую опцию



- Массив: перейдите на лист rlc_freq и выделите столбец с IPM.
- Номер_строки: поставьте курсор на поле "Номер..." в таблице corst_freq
- Номер_столбца: оставьте пустым, так как мы работаем с одним столбцом.



Нажмите OK.

Проверьте (с помощью поиска), что частота для данного слова указана правильно. Скопируем ячейку с формулой и вставим ее напротив всех остальных слов.

	A	B	C	D	E
1	Слово	Частота	IPM	Номер стр	IPM_rlc
2	в	35889	11520,56	3	37993,23
3	и	31435	10090,81	2	38334,7
4	на	13212	4241,124	5	13265,41
5	не	11811	3791,395	6	12860,17
6	с	11363	3647,585	8	9178,432
7	что	10265	3295,121	4	16305,4
8	как	7396	2374,156	9	8273,61
9	к	6790	2179,627	15	4742,013
10	а	5924	1901,636	20	3449,22
11	для	5652	1814,323	14	5009,075
12	по	5634	1808,545	16	4641,034
13	о	4912	1576,779	13	5022,361
14	из	4365	1401,189	17	4432,433

Теперь мы переставили все найденные данные из таблицы rlc_freq. Осталось добавить те слова из rlc_freq, которых не нашлось в основной таблице.

Шаг 5

Теперь все будет наоборот: добавим в таблицу rlc_freq столбец "Номер строки, на которой слово стоит в corst_freq".

Шаг 6

Для каждого слова в таблице rlc_freq найдем его позицию в таблице corst_freq и запишем в столбце "Номер строки..." (так же, как в Шаге 2).

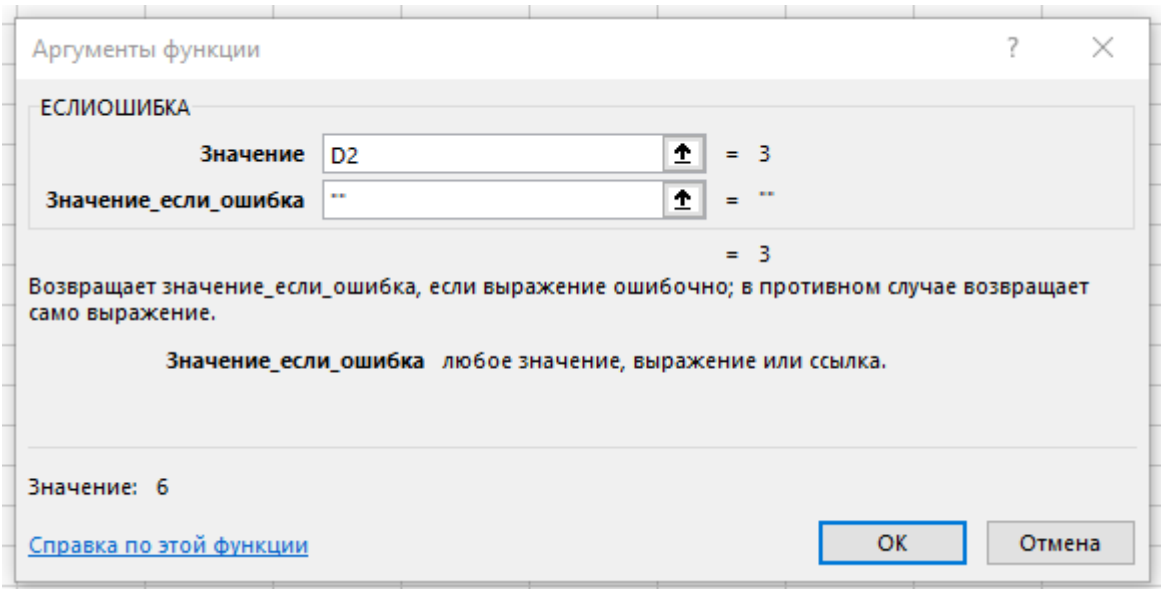
	A	B	C	D	E	F	G	H
1	Слово	Частота	IPM	Номер строки, на которой слово стоит в corst_freq				
2	и	28852	38334,7	3				
3	в	28595	37993,23	2				
4	что	12272	16305,4	7				
5	на	9984	13265,41	4				
6	не	9679	12860,17	5				
7	я	8489	11279,05	24				
8	с	6908	9178,432	6				
9	как	6227	8273,61	8				

Шаг 7

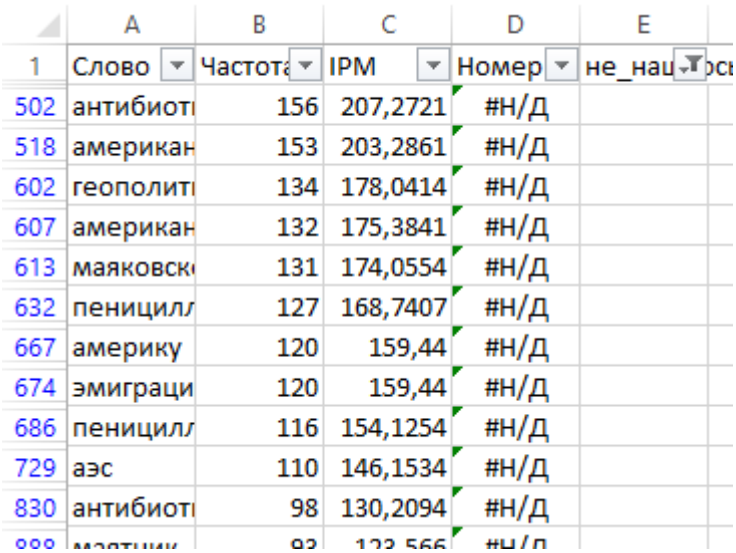
Теперь мы хотим взять из таблицы rlc_freq только те слова, которых не нашлось в таблице corst_freq. Для этого:

- добавим колонку "не_нашлось" в таблице rlc_freq
- На вкладке Формулы в разделе "Логические" найдем формулу "ЕСЛИОШИБКА", и в открывшемся мастере формулы:
- Значение: выбрать ячейку из столбца слева
- Значение_если_ошибка: пустая строка (две кавычки подряд, как в питоне)

(Комментарий: можно не использовать функцию ЕСЛИОШИБКА, а отфильтровать все ячейки "#Н/Д" в соответствующем столбце).



- Нажмите OK
- Скопируем ячейку с формулой и вставим ее напротив всех остальных слов.
- Создать фильтр в колонке "не_нашлось": оставить только те строки, в которых пустое значение.



	A	B	C	D	E
1	Слово	Частота	IPM	Номер	не_нац
502	антибиоти	156	207,2721	#Н/Д	
518	американ	153	203,2861	#Н/Д	
602	геополити	134	178,0414	#Н/Д	
607	американ	132	175,3841	#Н/Д	
613	маяковск	131	174,0554	#Н/Д	
632	пенициллі	127	168,7407	#Н/Д	
667	америку	120	159,44	#Н/Д	
674	эмиграци	120	159,44	#Н/Д	
686	пенициллі	116	154,1254	#Н/Д	
729	аэс	110	146,1534	#Н/Д	
830	антибиоти	98	130,2094	#Н/Д	

Добавим данные по этим словам в таблицу corst_freq (тут можно просто скопировать слова в столбец "Слова", а затем скопировать значения IPM в столбец "IPM_rlc"). *Примечание*: если вам не удается вставить скопировать слова на другую вкладку, значит, вы пытаетесь вставить содержание целого столца в часть другого столбца, что невозможно. Выделите именно диапазон ячеек от первого до последнего слова (и соответствующий диапазон частот).

Шаг 8

Теперь у нас в таблице corst_freq сведены все данные!

Давайте выделим нужные нам данные в отдельный лист. Создадим лист "merge", вставим туда колонки Слово, IPM_corst и IPM_rlc (используйте только вставку значений). В столбце IPM_rlc замените все вхождения #Н/Д на пустую строку. Укажите формат ячеек с IPM - числовой.

	A	B	C
1	Слово	IPM_corst	IPM_rlc
2	в	11520,56	37993,23
3	и	10090,81	38334,70
4	на	4241,12	13265,41
5	не	3791,40	12860,17
6	с	3647,58	9178,43
7	что	3295,12	16305,40
8	как	2374,16	8273,61
9	к	2179,63	4742,01
10	а	1901,64	3449,22
11	для	1814,32	5009,07

Отсортируйте таблицу по столбцу "Слово", а затем по столбцу "IPM_corst".

УРА!

Вопросы для проверки:

Ответы нужно писать в эту форму.

1. Какие слова довольно частотны в RLC, но не встречаются в CORST? (приведите 6 самых частотных слов)
2. Какие слова, наоборот, частотны в CORST, но не встречаются в RLC? (приведите 6 самых частотных слов)
3. Различаются ли первые 10 слов в таблице, отсортированной по столбцу "IPM_corst" по убыванию, и в таблице, отсортированной по столбцу "IPM_rlc" по убыванию?

Полезное

- частотный словарь НКРЯ (современный русский язык): http://dict.ruslang.ru/freq.php
- видовые пары глаголов по Грамматическому словарю А.А.Зализняка: таблица
- база данных синонимов русского языка: http://web-corpora.net/synonyms
- другие ресурсы школы лингвистики: http://web-corpora.net/ и https://linghub.ru/