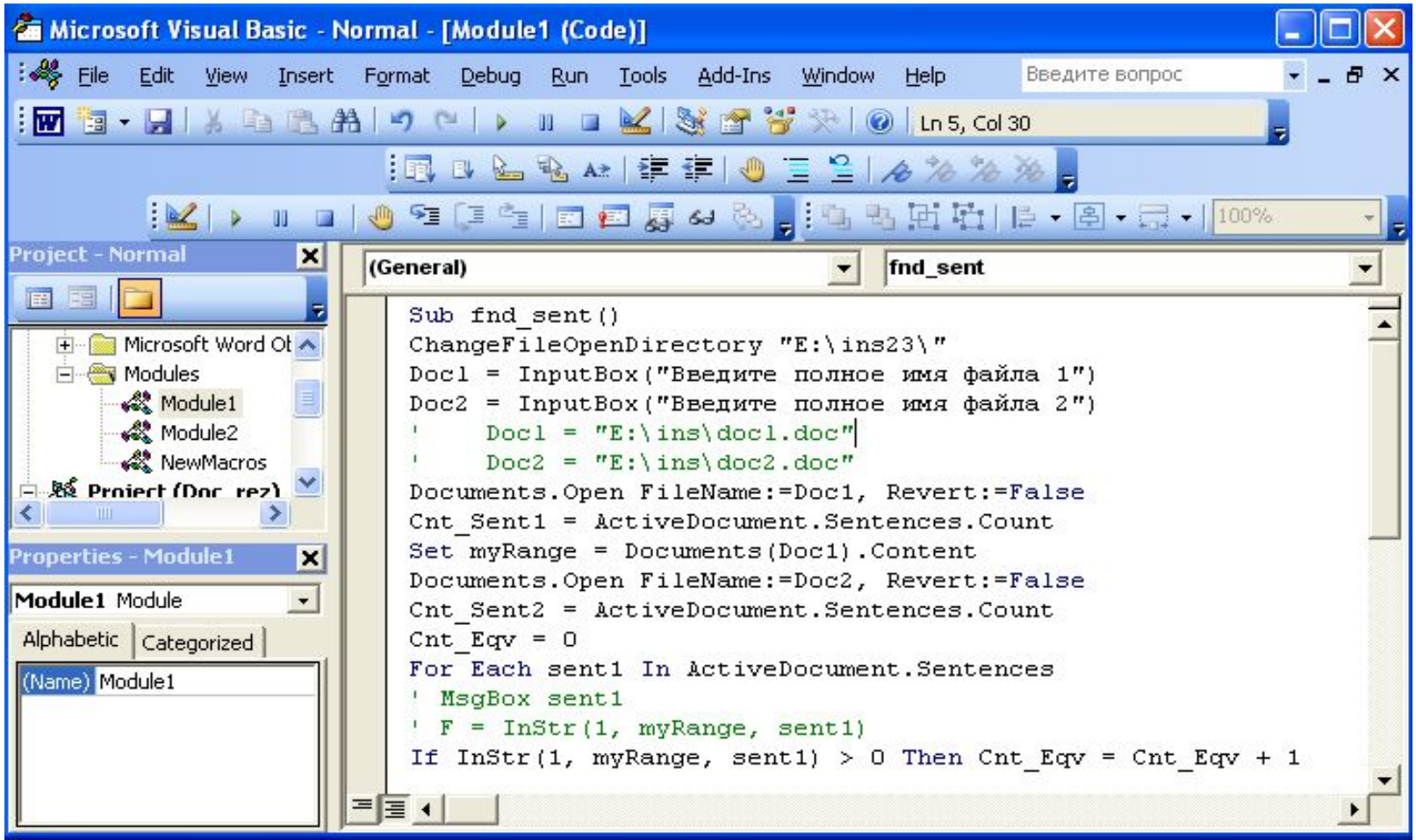
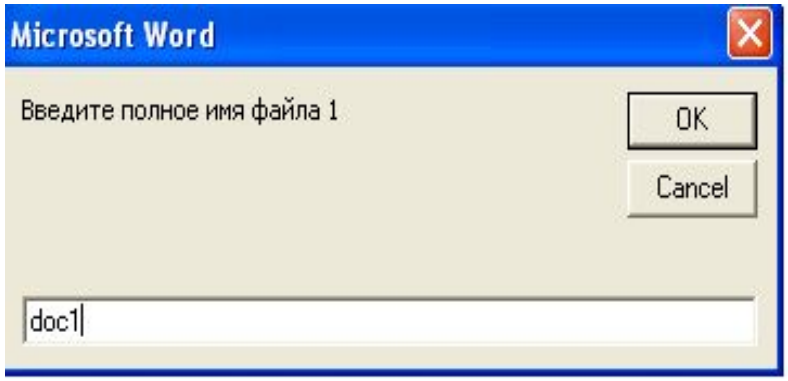
Тема работы:

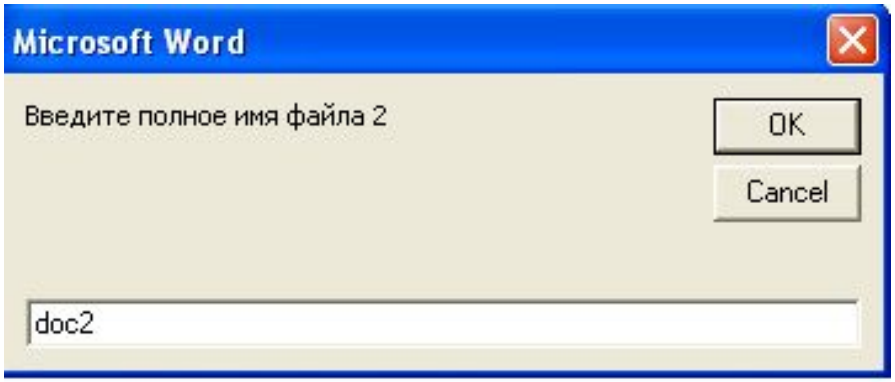
ИС лингвистического анализа заимствования в наборах текстовых файлов на примере пояснительных записок к дипломным работам.

Сысоева Е. гип-102

- Создаём в word макрос, который обрабатывает тексты пояснительных записок и сравнивает их между собой. Данный макрос выдаёт общее кол-во предложений, содержащихся в заданных пояснительных записках, также выдаёт кол-во одинаковых предложений, встречающихся и в обоих документах и подсчитывает % схожести документов, те на сколько % один документ похож на другой.







Выберем для примера главы 3 «экономические показатели» из 2-ух пояснительных записок. Сравнивая эти главы в двух разных пояснительных записках мы получим:

Результат работы макроса:

- doc1 256 предложений
- doc2 380 предложений
- 87 – общих предложений
- 33,984375 % процент схожести, т. е. насколько один документ похож на другой.

Сравним последовательно все главы записки 1 со всеми главами других пояснительных записок.

Записка1	Записка2	Записка3	Записка4	Записка5	Записка6
	Глава1	Глава1	Глава1	Глава1	Глава1
Глава1	Зап1 298 предл. Зап2 555 предл. 125 – общих предл. Процент схожести 41,9463087248322 %	Зап1 298 предл. Зап2 554 предл. 42 – общих предл. Процент схожести 14,0939597315436 %	Зап1 298 предл. Зап2 1640 предл. 10 – общих предл. Процент схожести 3,35570469798658 %	Зап1 298 предл. Зап2 1754 предл. 114 – общих предл. Процент схожести 38,255033557047 %	Зап1 298 предл. Зап2 855 предл. 129 – общих предл. Процент схожести 43,2885906040268 %
	Глава2	Глава2	Глава2	Глава2	Глава2
Глава2	Зап1 213 предл. Зап2 295 предл. 66 – общих предл. Процент схожести 30,9859154929577 %	Зап1 213 предл. Зап2 1029 предл. 25 – общих предл. Процент схожести 11,7370892018779 %	Зап1 213 предл. Зап2 751 предл. 12 - общих предл. Процент схожести 5,63380281690141 %	Зап1 213 предл. Зап2 305 предл. 30 – общих предл. Процент схожести 14,0845070422535 %	Зап1 213 предл. Зап2 855 предл. 31 – общих предл. Процент схожести 14,5539906103286 %
	Глава3	Глава3	Глава3	Глава3	Глава3
Глава3	Зап1 550 предл. Зап2 870 предл. 368 – общих предл. Процент схожести 66,9090909090909 %	Зап1 550 предл. Зап2 851 предл. 409 – общих предл. Процент схожести 74,3636363636364 %	Зап1 550 предл. Зап2 634 предл. 269 – общих предл. Процент схожести 48,9090909090909 %	Зап1 550 предл. Зап2 495 предл. 247 – общих предл. Процент схожести 44,9090909090909 %	Зап1 550 предл. Зап2 385 предл. 149 – общих предл. Процент схожести 27,0909090909091 %
	Глава4	Глава4	Глава4	Глава4	Глава4
Глава4	Зап1 365 предл. Зап2 159 предл. 33 – общих предл. Процент схожести 9,04109589041096 %	Зап1 365 предл. Зап2 256 предл. 35 – общих предл. Процент схожести 9,58904109589041 %	Зап1 365 предл. Зап2 267 предл. 55 – общих предл. Процент схожести 15,0684931506849 %	Зап1 365 предл. Зап2 262 предл. 48 – общих предл. Процент схожести 13,1506849315068 %	Зап1 365 предл. Зап2 380 предл. 89 – общих предл. Процент схожести 24,3835616438356 %

Аналогично все главы записки 2 сравним со всеми главами других пояснительных записок.

Записка2	Записка3	Записка4	Записка5	Записка6
	Глава1	Глава1	Глава1	Глава1
Глава1	doc1 555 doc2 554 85 15,3153153153153	doc1 555 doc2 1640 92 16,5765765765766	doc1 555 doc2 1754 661 119,099099099099	doc1 555 doc2 855 141 25,4054054054054
	Глава2	Глава2	Глава2	Глава2
Глава2	doc1 295 doc2 1029 24 8,13559322033898	doc1 295 doc2 751 14 4,74576271186441	doc1 295 doc2 305 22 7,45762711864407	doc1 295 doc2 276 41 13,8983050847458
	Глава3	Глава3	Глава3	Глава3
Глава3	doc1 870 doc2 851 437 50,2298850574713	doc1 870 doc2 634 226 25,9770114942529	doc1 870 doc2 495 274 31,4942528735632	doc1 870 doc2 385 129 14,8275862068966
	Глава4	Глава4	Глава4	Глава4
Глава4	doc1 159 doc2 256 28 17,6100628930818	doc1 159 doc2 267 57 35,8490566037736	doc1 159 doc2 262 90 56,6037735849057	doc1 159 doc2 380 120 75,4716981132076

И все главы записки 3 со всеми главами других пояснительных записок.

Записка3	Записка4	Записка5	Записка6
	Глава1	Глава1	Глава1
Глава1	doc1 554 doc2 1640 107 19,3140794223827	doc1 554 doc2 1754 126 22,7436823104693	doc1 554 doc2 855 71 12,8158844765343
	Глава2	Глава2	Глава2
Глава2	doc1 1029 doc2 751 199 19,3391642371234	doc1 1029 doc2 305 62 6,02526724975705	doc1 1029 doc2 276 35 3,40136054421769
	Глава3	Глава3	Глава3
Глава3	doc1 851 doc2 634 242 28,4371327849589	doc1 851 doc2 495 252 29,6122209165687	doc1 851 doc2 385 164 19,2714453584019
	Глава4	Глава4	Глава4
Глава4	doc1 256 doc2 267 42 16,40625	doc1 256 doc2 262 46 17,96875	doc1 256 doc2 380 87 33,984375

Проанализируем полученные результаты. Сравним 3 дипломных работы.

			Записка 1	Записка2	Записка 3	Записка 1	Записка 2	Записка 3	Записка 1	Записка 2	Записка 3	Записка 1	Записка 2	Записка 3
			Гл1	Гл1	Гл1	Гл2	Гл2	Гл2	Гл3	Гл3	Гл3	Гл4	Гл4	Гл4
Записка1	Гл1	298	-	41,9%	14,1%									
Записка1	Гл2	213				-	31,00%	11,70%						
Записка1	Гл3	550							-	66,90%	74,40%			
Записка1	Гл4	365										-	9,00%	9,60%
Записка2	Гл1	555	22,50%	-	15,30%									
Записка2	Гл2	295				22,40%	-	8,10%						
Записка2	Гл3	870							42,30%	-	50,20%			
Записка2	Гл4	159										20,80%	-	17,60%
Записка3	Гл1	554	8%	15%	-									
Записка3	Гл2	1029				2%	2%	-						
Записка3	Гл3	851							48%	51%	-			
Записка3	Гл4	256										14%	11%	-

Например, 14,1 % показывает, что гл. 1 записки 1 явл-ся наименее схожей с гл.1 записки 3. Это минимальный процент схожести по данной главе. 41,9 % показывает, что гл.1 записки 1 явл-ся максимально схожей с гл.1 записки 2. 28,0 % - это средняя схожесть гл. 1 док -та 1 с гл. 1 других документов.

min	сред	max
14,1%	28,0%	41,9%
11,7%	21,4%	31,0%
66,9%	70,7%	74,4%
9,0%	9,3%	9,6%
15,3%	15,3%	22,5%
8,1%	15,3%	22,4%
42,3%	46,3%	50,2%
17,6%	19,2%	20,8%
8,0%	15,0%	15,0%
2,0%	2,0%	2,0%
48,0%	49,5%	51,0%
11,0%	12,5%	14,0%

Глава1: Системно-техническая часть.
Глава2: Конструкторско-технологическая часть.
Глава3: Экономические показатели.
Глава4: Безопасность.

	min	сред	Средне взвешен ное	max
Записка1	9,0%	32,3%	38,7%	74,4%
Записка2	8,1%	24,9%	30,0%	50,2%
Записка3	2,0%	18,9%	20,7%	51,0%

9,0 % - из 4-ёх глав записки 1 минимально схожи гл. 4 записки 1 и записки 2.
74,4 % - из 4-ёх глав записки 1 максимально схожи гл.3 записки 1 и записки 3.
32,3 % - средний % схожести всех глав записки 1 со всеми главами других записок.
38,7 % - средневзвешенный % схожести при учёте того, что главы занимают неравные части в дипломе.