НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Научно-учебная лаборатория «Разработка программного обеспечения для проведения корпусных исследований английского языка»

Подсистема визуализации статистической информации для системы лингвистических исследований

Автор: Алексеева Анастасия

Пермь, 2020

Содержание

Система лингвистических исследований

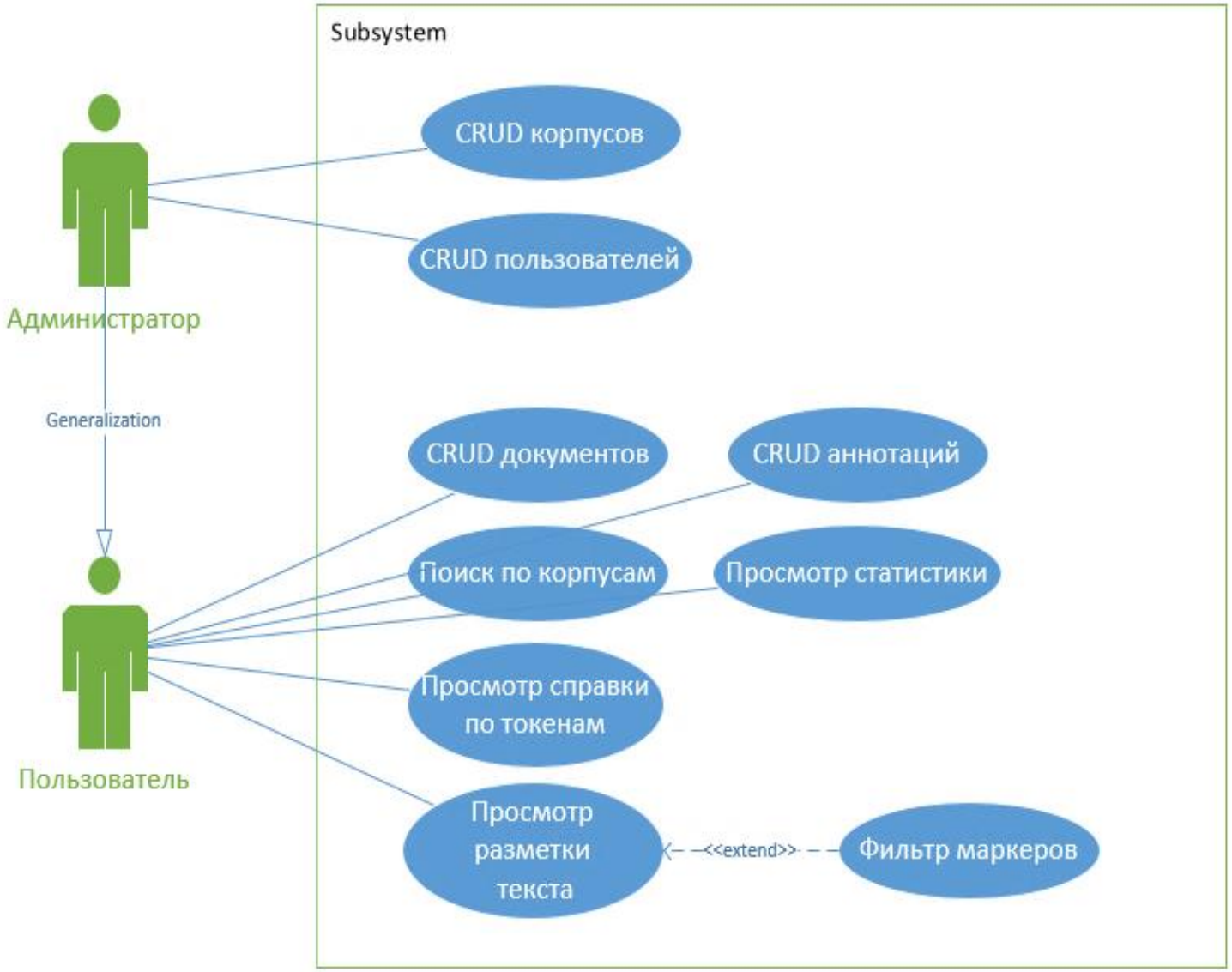
Корпусная лингвистика - раздел языкознания, занимающийся изучением языка через анализ лингвистических корпусов, т.е. собраний электронных текстов.

В НИУ ВШЭ разрабатывается система для исследований в области корпусной лингвистики

Система собирает корпусы текстов, написанных на английском языке

В качестве корпусов используются работы студентов и работы экспертов из тех же предметных областей

Система лингвистических исследований



Система лингвистических исследований

Система лингвистических исследований

Работы студентов оцениваются по двум критериям

Качественный критерий оценивает соответствие академическому стилю текста

Количественный критерий проверяет наличие в тексте слов, соответствующих маркерам академического текста

В процессе работы система собирает большое количество статистики по корпусам

Отчеты по корпусам получаются громоздкими и сложными для восприятия и сравнения

Для упрощения восприятия статистики необходимо визуализировать информацию – представить ее в виде графиков и диаграмм

Постановка задачи

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1		Abstract semantic verbs	Abstract semantic verbs	Abstract suffix nouns	Abstract suffix nouns	Archaisms	Archaisms	Complex conjunctions	Complex conjunctions	Complex prepositions	Complex prepositions	Desemantisized verbs	Desemantisized verbs	Future	Future	Intensifying adverbs	Intensifying adverbs	IPronoun	IPronoun	Noun	Noun	OrSuffix	OrSuffix
2	Среднее	0,004702	0,005451	0,010716	0,01139	0,00032	0,00028	0,002672	0,00181	0,001254	0,0012	0,003536	0,000373	0,001195	0,00231	0,0008	0,00052	0,001391	0,00057	0,048709	0,0519	0,00555139	0,00137
3	Ст. откл.	0,001569	0,001056	0,002118	0,00299	0,000476	0,00046	0,001174	0,00105	0,000242	0,00036	0,000814	0,00103	0,000887	0,00109	0,000555	0,00057	0,001202	0,00093	0,015501	0,00405	0,01594721	0,00077
4	0	2,846854	0,000619	0,000518	0,094223	668,5237	720,6044	25,53646	85,99309	0,00253	4,284108	0,039379	362,7401	181,4929	38,74473	254,2852	461,6503	169,8296	355,5132	0,184624	2,16E-34	23,5456845	106,4173
5	0,0005	7,033816	0,00637	0,001664	0,175687	780,1479	773,5411	61,44406	174,4695	13,0205	167,3414	0,469307	384,3895	331,0129	92,19677	621,2227	699,468	252,108	427,7568	0,204214	1,04E-33	23,7923834	273,6578
6	0,001	15,70014	0,052425	0,005055	0,318549	302,1592	254,7781	123,3305	282,1604	950,3343	949,7008	3,836095	321,8151	439,244	177,7602	673,7725	490,9612	314,7767	385,4828	0,225647	4,95E-33	24,0180448	461,6154
7	0,0015	31,65939	0,344795	0,014523	0,561653	38,84128	25,74737	206,5059	363,7416	983,5491	783,0885	21,50633	212,863	424,0747	277,6957	324,4286	159,6434	330,5684	260,1836	0,24907	2,32E-32	24,2220236	510,7752
8	0,002	57,67509	1,812263	0,039467	0,962976	1,657104	0,798354	288,4469	373,7753	14,434	93,8157	82,69667	111,2379	297,8891	351,4953	69,35293	24,04801	291,9866	131,529	0,274638	1,07E-31	24,4037331	370,7288
9	0,0025	94,92065	7,612404	0,101434	1,605528	0,023464	0,007595	336,1016	306,1601	0,003004	1,63298	218,0994	45,92663	152,2448	360,4837	6,581885	1,678154	216,9238	49,80011	0,302517	4,86E-31	24,5626478	176,5062
10	0,003	141,1301	25,55423	0,246562	2,603012	0,00011	2,22E-05	326,6981	199,8974	8,86E-09	0,00413	394,5171	14,98081	56,6117	299,5487	0,277316	0,054251	135,5484	14,12231	0,332879	2,17E-30	24,6983061	55,12396
11	0,0035	189,5682	68,55574	0,566836	4,103832	1,72E-07	1,99E-08	264,9074	104,0366	3,71E-16	1,52E-06	489,4653	3,860689	15,31601	201,6811	0,005187	0,000812	71,24006	2,999494	0,365907	9,57E-30	24,8103121	11,2927
12	0,004	230,037	146,9823	1,232474	6,291563	8,9E-11	5,46E-12	179,1896	43,16045	2,2E-25	8,1E-11	416,5074	0,786056	3,014814	110,0217	4,31E-05	5,64E-06	31,49176	0,477153	0,401794	4,15E-29	24,8983379	1,517513
13	0,0045	252,1835	251,8412	2,534473	9,37957	1,53E-14	4,6E-16	101,1121	14,2727	1,85E-36	6,28E-16	243,0907	0,126445	0,431769	48,63038	1,59E-07	1,81E-08	11,7088	0,05685	0,440742	1,78E-28	24,9621253	0,133765
14	0,005	249,7596	344,8486	4,92932	13,59761	8,72E-19	1,19E-20	47,59535	3,762243	2,21E-49	7,08E-22	97,31032	0,01607	0,04499	17,41616	2,6E-10	2,7E-11	3,661599	0,005073	0,482962	7,47E-28	25,0014866	0,007735
15	0,0055	223,4675	377,3727	9,067256	19,16894	1,65E-23	9,45E-26	18,68949	0,790511	3,74E-64	1,16E-28	26,7174	0,001614	0,003411	5,053743	1,89E-13	1,86E-14	0,963101	0,000339	0,528677	3,1E-27	25,0163058	0,000293
16	0,006	180,6314	330,0295	15,77445	26,27778	1,04E-28	2,3E-31	6,122117	0,132401	8,96E-81	2,76E-36	5,03123	0,000128	0,000188	1,1882	6,09E-17	5,95E-18	0,213067	1,7E-05	0,578116	1,27E-26	25,0065393	7,3E-06
17	0,0065	131,9042	230,6615	25,95505	35,02957	2,16E-34	1,72E-37	1,672928	0,017676	3E-99	9,53E-45	0,649829	8,02E-06	7,55E-06	0,226351	8,73E-21	8,8E-22	0,039646	6,36E-07	0,631522	5,09E-26	24,972216	1,19E-07
18	0,007	87,01836	128,8362	40,39048	45,40842	1,5E-40	3,94E-44	0,38135	0,001881	1,5E-119	4,79E-54	0,057566	3,97E-07	2,2E-07	0,034937	5,55E-25	6,04E-26	0,006205	1,79E-08	0,689144	2,01E-25	24,9134368	1,28E-09
19	0,0075	51,86204	57,50964	59,4464	57,23917	3,43E-47	2,77E-51	0,072517	0,00016	1E-141	3,49E-64	0,003498	1,55E-08	4,68E-09	0,004369	1,57E-29	1,92E-30	0,000817	3,76E-10	0,751241	7,86E-25	24,8303747	8,95E-12
20	0,008	27,92383	20,51558	82,74878	70,16262	2,62E-54	5,99E-59	0,011504	1,08E-05	9,7E-166	3,7E-75	0,000146	4,8E-10	7,24E-11	0,000443	1,96E-34	2,82E-35	9,04E-05	5,92E-12	0,818082	3,02E-24	24,7232737	4,12E-14
21	0,0085	13,58273	5,848804	108,9399	83,63222	6,62E-62	3,97E-67	0,001522	5,82E-07	1,3E-191	5,7E-87	4,17E-06	1,17E-11	8,14E-13	3,64E-05	1,09E-39	1,92E-40	8,42E-06	6,98E-14	0,889944	1,14E-23	24,5924474	1,25E-16
22	0,009	5,968782	1,332572	135,6444	96,93864	5,56E-70	8,06E-76	0,000168	2,5E-08	2,6E-219	1,3E-99	8,17E-08	2,26E-13	6,66E-15	2,42E-06	2,7E-45	6,08E-46	6,6E-07	6,17E-16	0,967112	4,25E-23	24,4382778	2,47E-19
23	0,0095	2,369578	0,242636	159,7372	109,2636	1,55E-78	5,03E-85	1,55E-05	8,56E-10	7,2E-249	4,2E-113	1,1E-09	3,45E-15	3,97E-17	1,3E-07	2,96E-51	8,89E-52	4,35E-08	4,08E-18	1,049878	1,56E-22	24,2612132	3,21E-22
24	0,01	0,849851	0,035307	177,9098	119,7595	1,43E-87	9,61E-95	1,19E-06	2,34E-11	2,8E-280	2E-127	1,01E-11	4,15E-17	1,72E-19	5,69E-09	1,44E-57	6,03E-58	2,41E-09	2,02E-20	1,138541	5,64E-22	24,0617662	2,74E-25
25	0,0105	0,27536	0,004106	187,4058	127,6437	4,4E-97	5,6E-105	7,62E-08	5,08E-13	0	1,3E-142	6,41E-14	3,95E-19	5,41E-22	2,01E-10	3,11E-64	1,89E-64	1,12E-10	7,52E-23	1,233409	2,01E-21	23,8405112	1,53E-28
26	0,011	0,080602	0,000382	186,7049	132,2953	4,5E-107	1E-115	4,07E-09	8,82E-15	0	1,3E-158	2,78E-16	2,97E-21	1,24E-24	5,77E-12	2,99E-71	2,75E-71	4,4E-12	2,09E-25	1,334791	7,04E-21	23,5980815	5,62E-32
27	0,0115	0,021315	2,83E-05	175,921	133,3352	1,5E-117	5,6E-127	1,82E-10	1,22E-16	0	1,9E-175	8,27E-19	1,76E-23	2,07E-27	1,34E-13	1,27E-78	1,86E-78	1,45E-13	4,35E-28	1,443005	2,43E-20	23,3351663	1,35E-35
28	0,012	0,005092	1,68E-06	156,7722	130,6775	1,7E-128	9,5E-139	6,76E-12	1,34E-18	0	4,1E-193	1,69E-21	8,27E-26	2,51E-30	2,52E-15	2,41E-86	5,79E-86	4,02E-15	6,79E-31	1,558369	8,27E-20	23,0525077	2,14E-39
29	0,0125	0,001099	7,98E-08	132,1326	124,541	6,3E-140	5E-151	2,1E-13	1,18E-20	0	1,3E-211	2,36E-24	3,07E-28	2,22E-33	3,85E-17	2,02E-94	8,38E-94	9,38E-17	7,93E-34	1,681206	2,77E-19	22,750897	2,21E-43
30	0,013	0,000214	3,03E-09	105,3272	115,4195	7,8E-152	7,9E-164	5,43E-15	8,27E-23	0	5,6E-231	2,27E-27	8,98E-31	1,43E-36	4,76E-19	7,5E-103	5,6E-102	1,84E-18	6,94E-37	1,81184	9,13E-19	22,4311708	1,5E-47
31	0,0135	3,77E-05	9,16E-11	79,40724	104,0163	3,2E-164	3,9E-177	1,17E-16	4,61E-25	0	3,6E-251	1,49E-30	2,08E-33	6,67E-40	4,76E-21	1,2E-111	1,7E-110	3,04E-20	4,55E-40	1,950593	2,97E-18	22,0942077	6,7E-52

Лист1

Пример исходных данных

Постановка задачи

Для визуализации пользователи системы используют программы Microsoft Office

Но визуализировать всю статистику очень долго и сложно

Более сложные приложения для визуализации требуют специальных знаний для работы

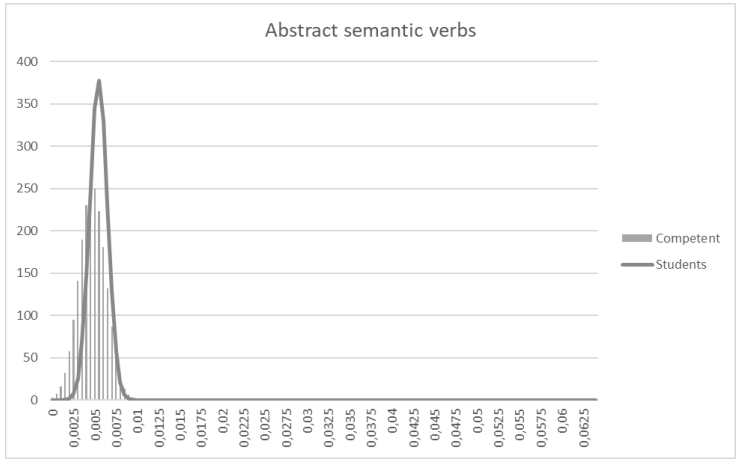
Поэтому цель данной работы – создание подсистемы визуализации для системы лингвистических исследований

Анализ предметной области

Для каждого маркера рассчитываются статистические показатели:

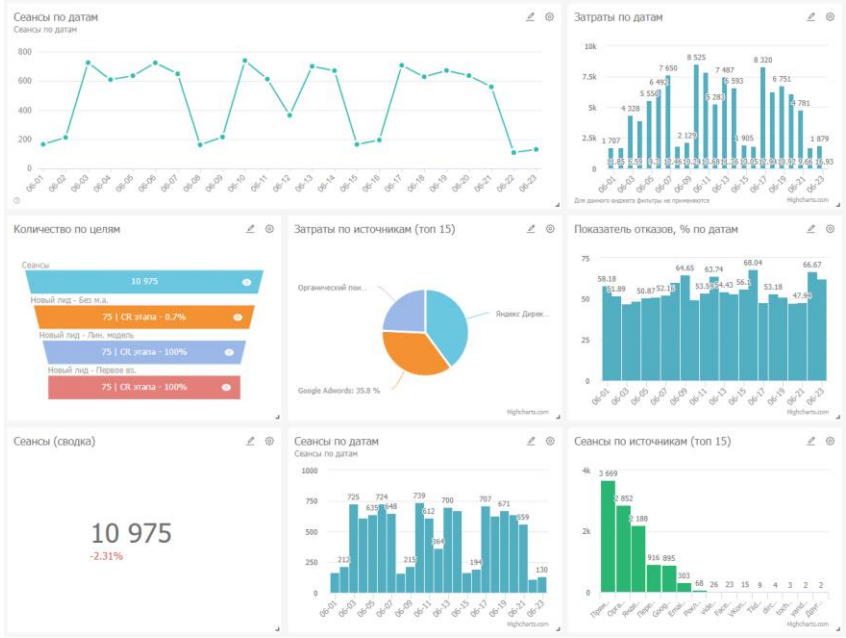
- Мода
- Медиана
- Среднее значение на 1000 слов
- Количество употреблений во всем корпусе

Аспект	Маркер	Характеристики маркера, примеры
Nominalization	Noun	
	Noun with abstract suffix	Суффиксы: -ment, -ion, -ation, -ition, -tion, -sion, -f, -ness, -ce, -cy, -ity, -dom, -th, -ery, -ry, -ise, -ice, -hood, -ics, -ship
	Noun with -or suffix	-
Personal Stance	Pronoun I	-
	Pronoun we	-
	Pronoun he/she	-
	That of/those of	-
Verb	Desemanticised verbs	Be, become, seem, remain, grow, consider
	Verbs of broad abstract semantics	Be, exist, have, appear, occur, alter, continue, contribute, discuss, involve, investigate, conduct, consider, illustrate, assume, find, calculate, demonstrate, identify, analyse, support, challenge, examine, affect, provide, include, classify, establish
	Future tense	-
	Past tense	-
	Present tense	-
	Passive voice	-
Adverbs	Intensifying adverbs	Clearly, dramatically, completely, considerably, essentially, significantly, markedly, perfectly
Attributes	Prepositive attributes	Характерны для технических текстов
	Postpositive attributes	-
Cohesiveness	Complex conjunctions	Not merely, but also, both and, as ... as, neither ... nor, the ... the, not so ... as
	Archaisms	Thereby, therewith, hereby
	Complex prepositions	Throughout, within, in accordance with, instead of, according to, because of, due to, regardless of
	Logic connectors	Since, therefore, (it) follow\s (that), so, thus, lead\s to, result\s in



Примеры диаграмм, которые строятся по статистическим показателям

Для представления диаграмм в подсистеме визуализации оптимальнее всего использовать интерфейс дашборд



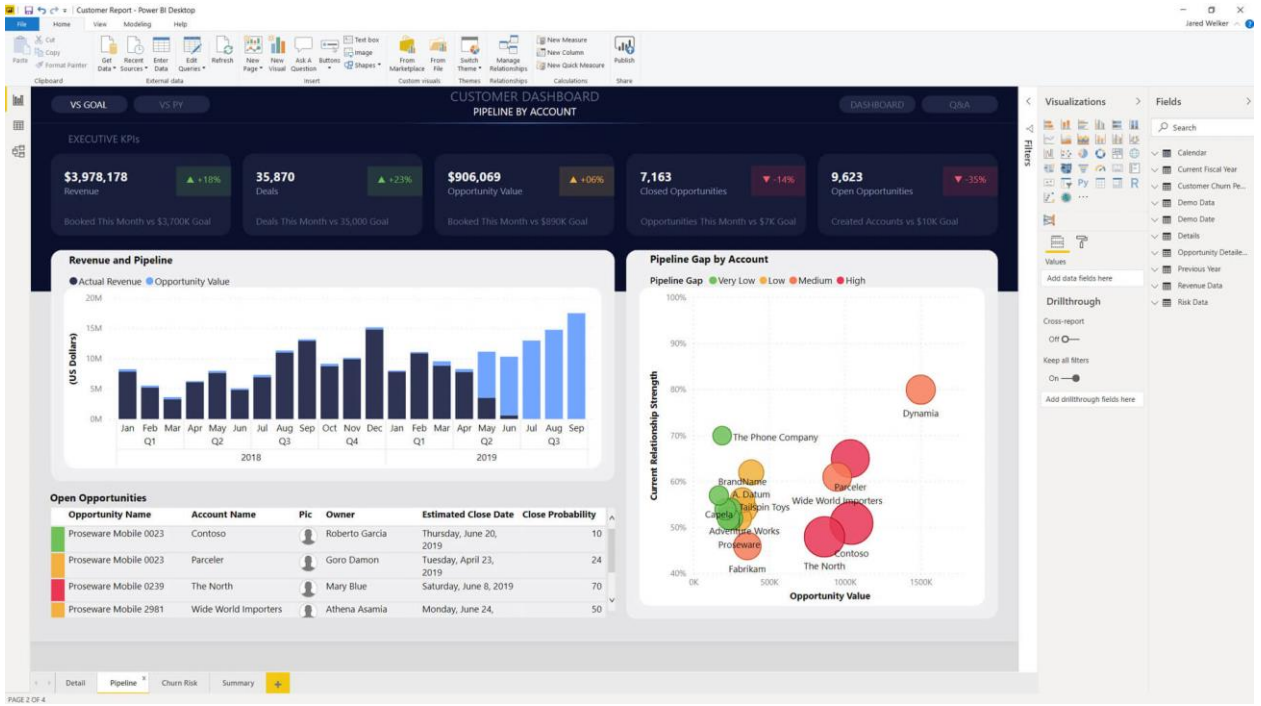
Пример дашборда

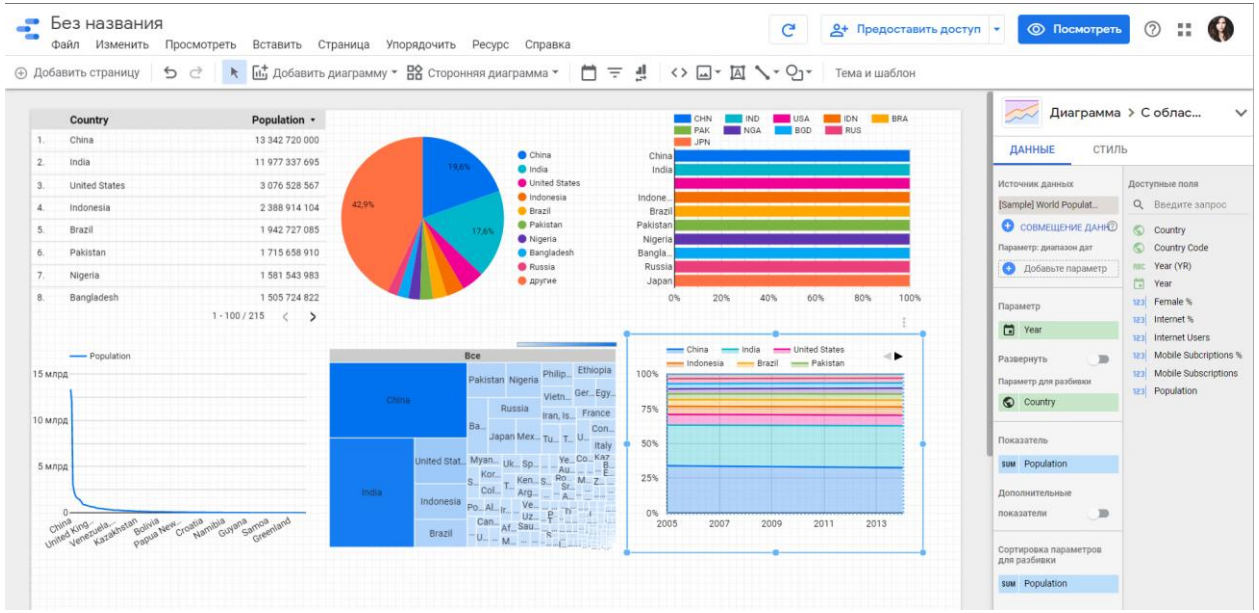
Анализ предметной области

Подсистема визуализации статистической информации для системы лингвистических исследований

Примеры программ визуализации данных

Microsoft Power BI

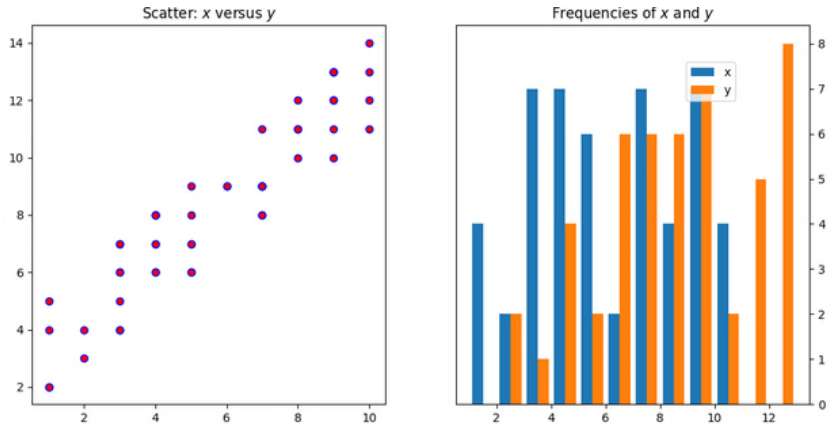




Google Data Studio

№	Название	Язык	Краткое описание возможностей
1	Matplotlib	Python	+ поддерживает множество типов графиков и диаграмм; диаграммы можно настраивать; поддерживает множество форматов изображений - диаграммы очень однообразны; высокая сложность настройки.
2	Plotly	Python	+ красочные и интерактивные диаграммы; большой выбор типов диаграмм; подходит для работы с web-приложениями - настройка и создание занимают много времени.
3	Chart.js	JavaScript	+ подходит для работы с web-приложениями; небольшой размер и возможность работы только с необходимыми модулями +- создает ограниченное количество диаграмм и графиков, но все они хорошо настраиваются.
4	Dygraphs	JavaScript	+ достаточно простое использование; может работать с большими наборами данных; подходит для работы с web-приложениями; графики масштабируются, умеют «подсказывать» пользователю.
5	D3.js	JavaScript	+ не просто библиотека, а полноценный фреймворк, содержащий набор инструментов для визуализации данных; - сложность в освоении из-за обилия данных; каждый график необходимо прорисовывать вручную.
6	JavaScript InfoVis Toolkit	JavaScript	+- содержит в себе не часто используемые виды визуализации данных, поэтому предназначен для решения особых типов задач.
7	jpGraph	PHP	+ легкий инструмент для создания простых типов графиков на PHP; обширная документация; широкий выбор диаграмм.
8	Google Charts	JavaScript	+ помимо стандартных диаграмм позволяет создавать деревья, интерактивные карты, датчики и другие виды визуализации; удобная и обширная документация с примерами.
9	jQuery Flot	JavaScript	+ большой выбор графиков; простая в использовании.

Выбор инструмента визуализации: обзор библиотек



Matplotlib

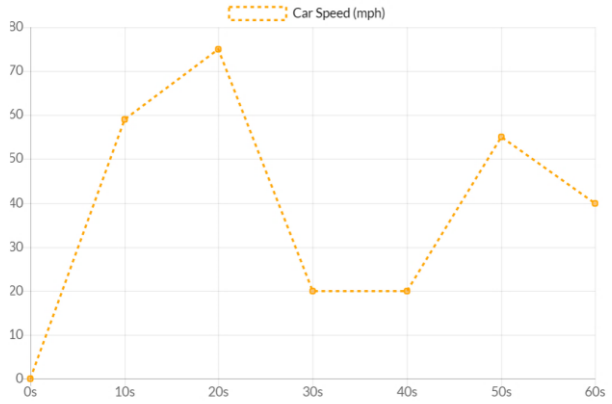
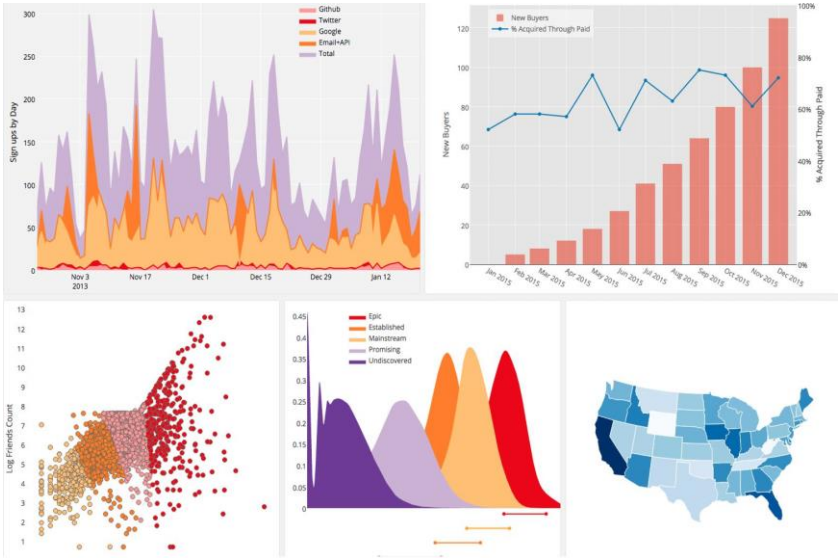
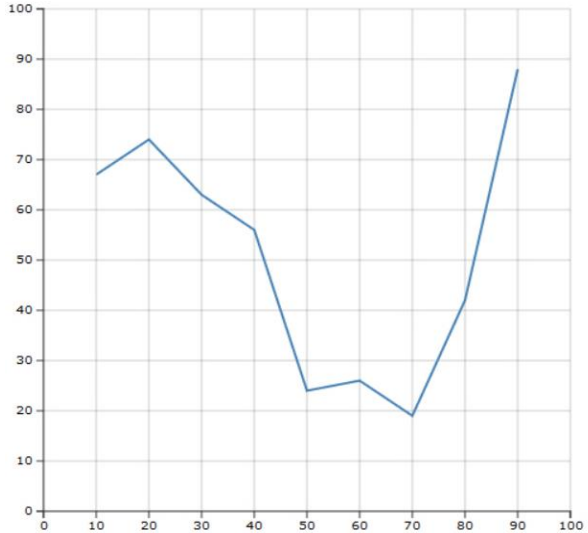


Chart.js

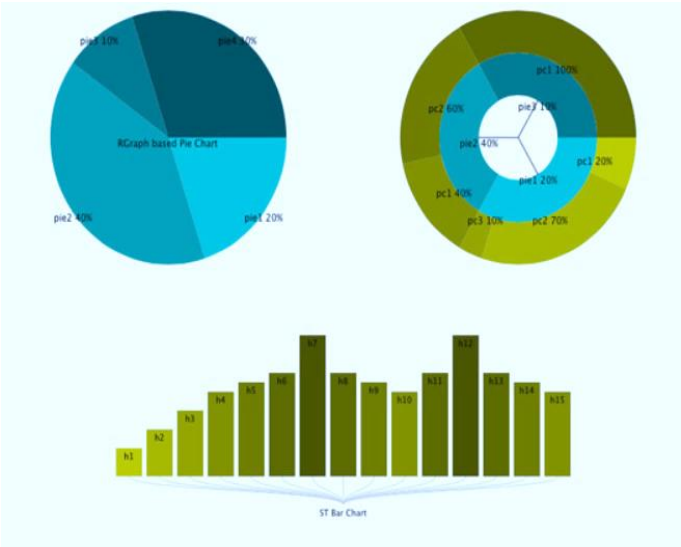
Dygraphs



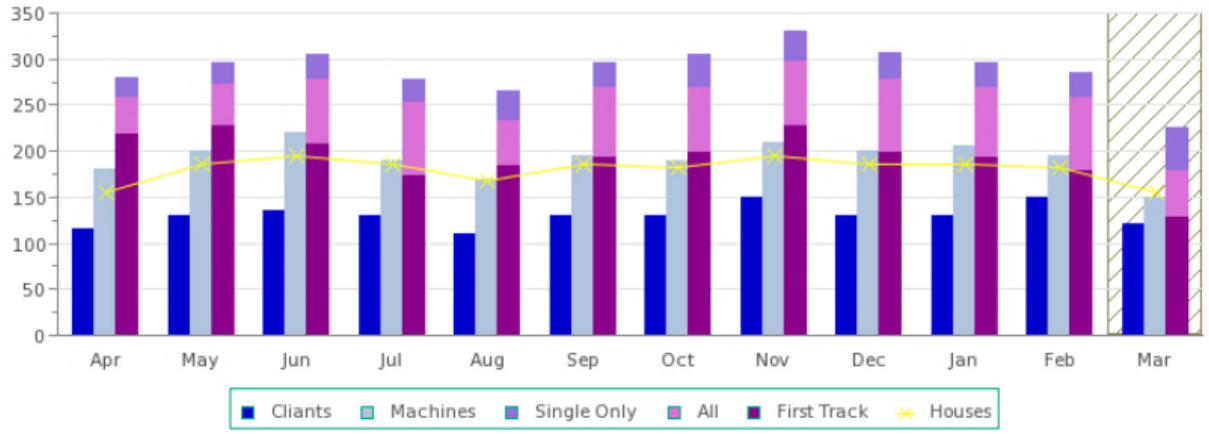
Plotly



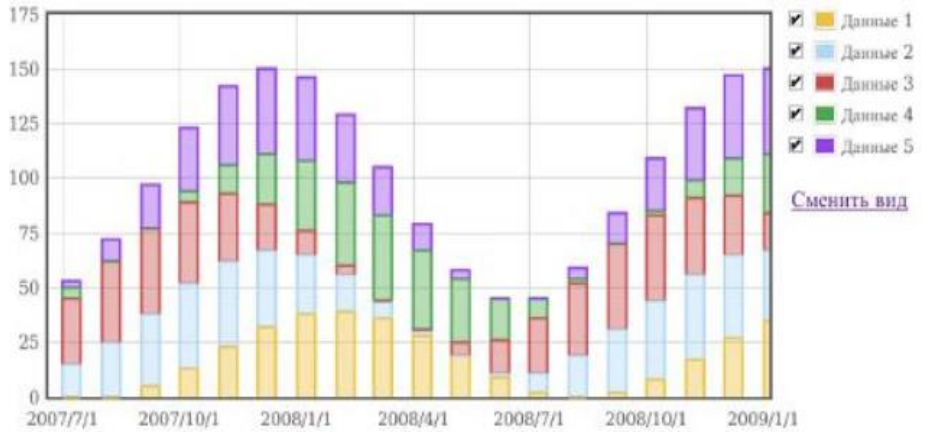
D3.js



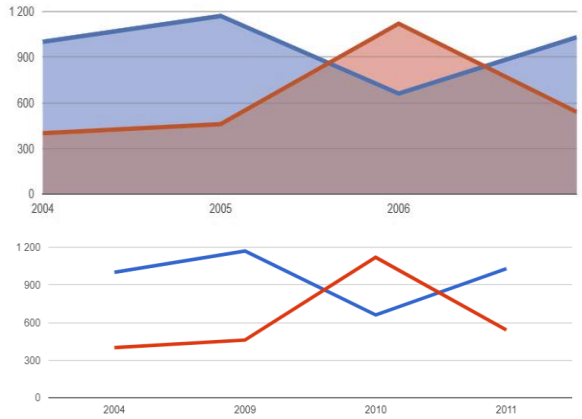
InfoVis Toolkit

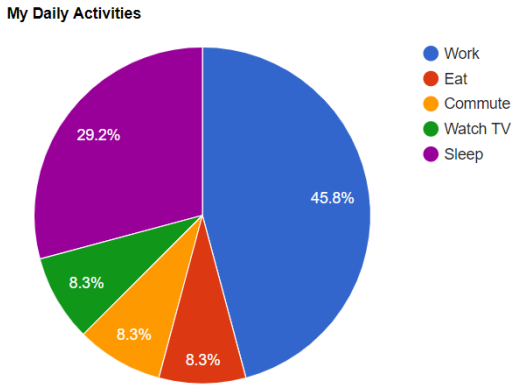


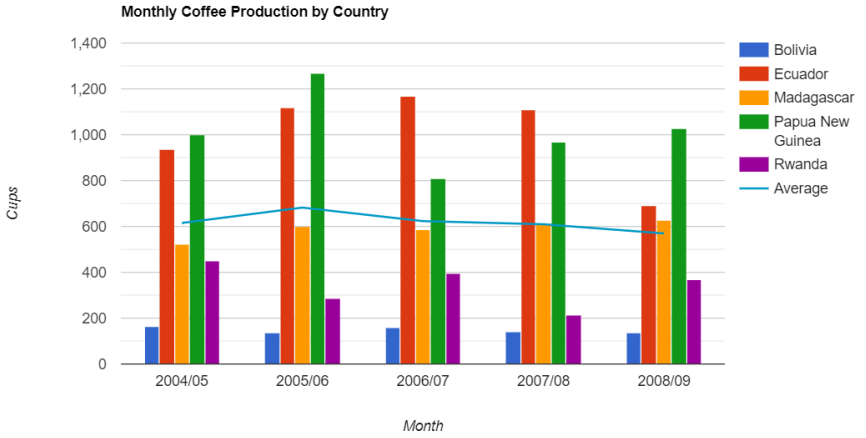
jpGraph



jQuery Flot







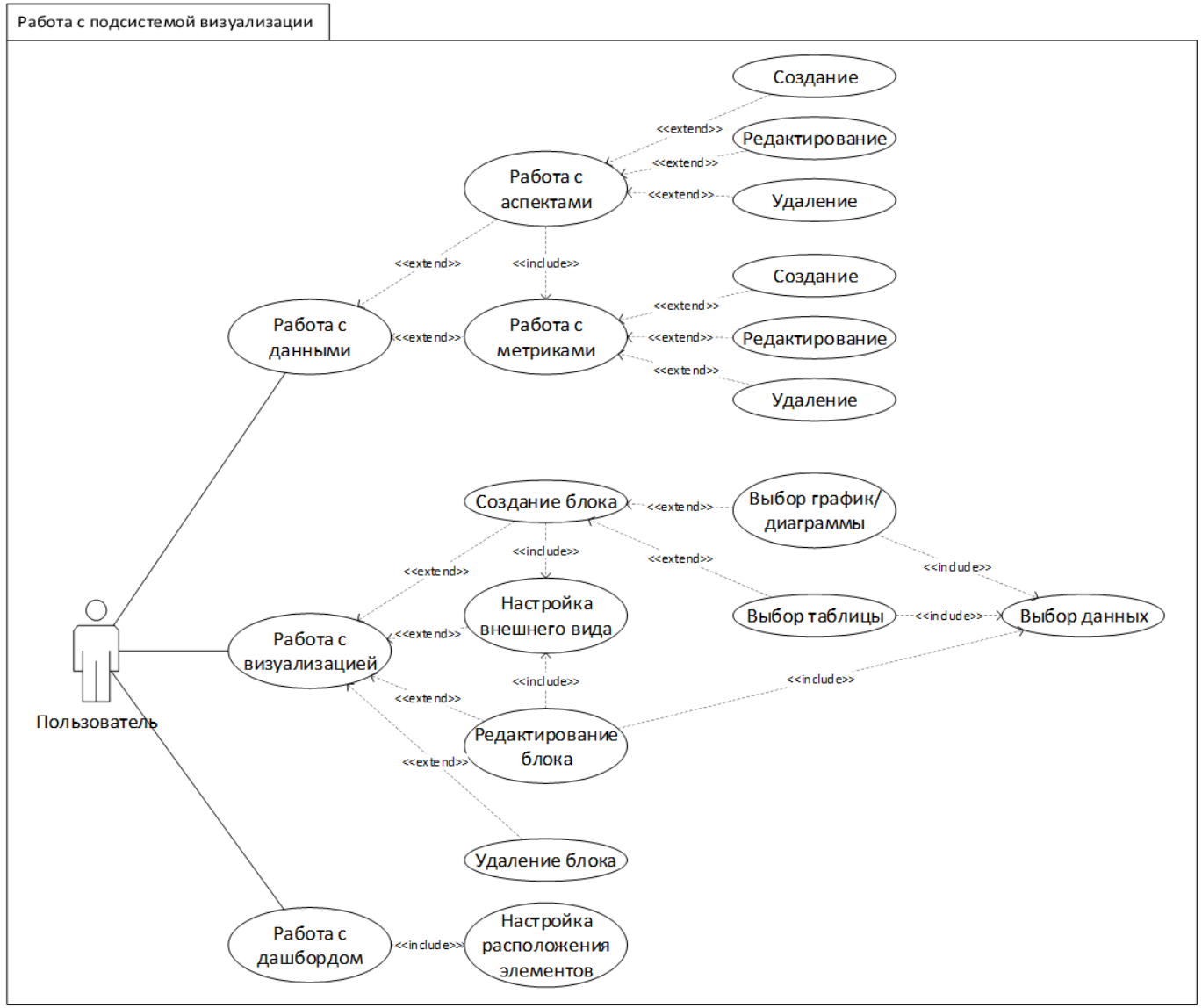
Google Charts

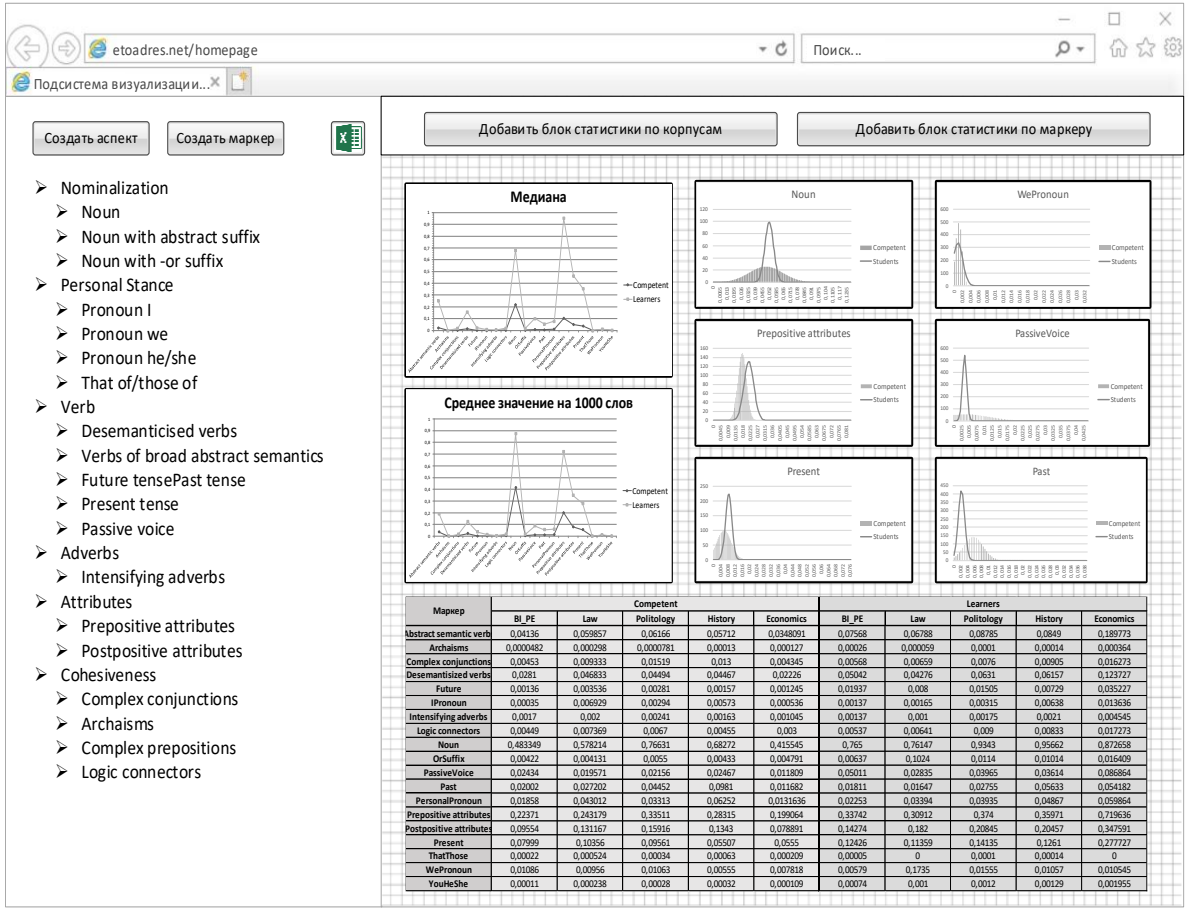
Название	Наличие нужных диаграмм	Наличие документации	Красочность и интерактивность	Легкость в использовании	Легкость в освоении	Итог
Matplotlib	10	9	3	3	3	28
Plotly	10	9	10	7	7	43
Chart.js	5	9	8	9	9	40
Dygraphs	10	9	7	9	9	44
D3.js	5	8	3	2	2	20
JavaScript InfoVis Toolkit	5	5	7	5	5	27
jpGraph	10	9	8	6	5	38
Google Charts	**10**	**10**	**10**	**9**	**9**	**48**
jQuery Flot	10	9	8	8	8	43

Выбор инструмента визуализации: оценка библиотек

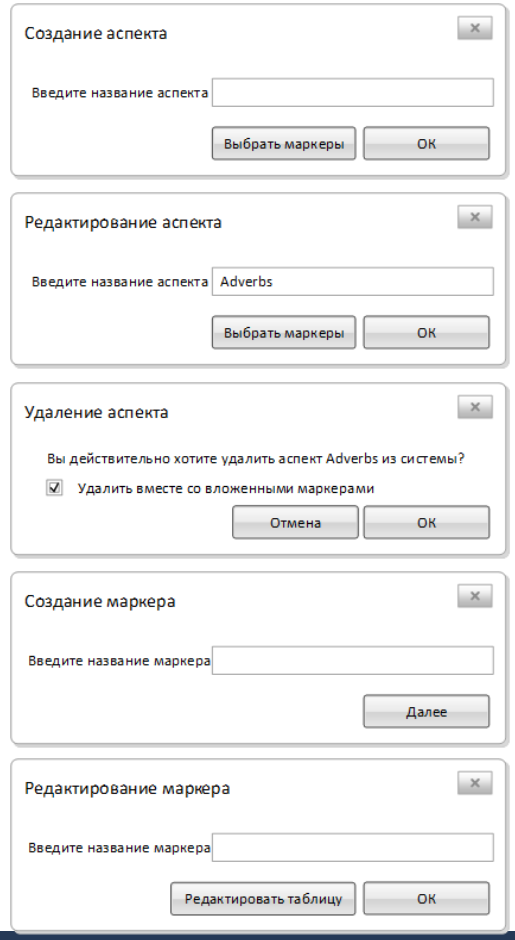
от **0** – признак отсутствует или выражен очень слабо до **10** – признак сильно выражен

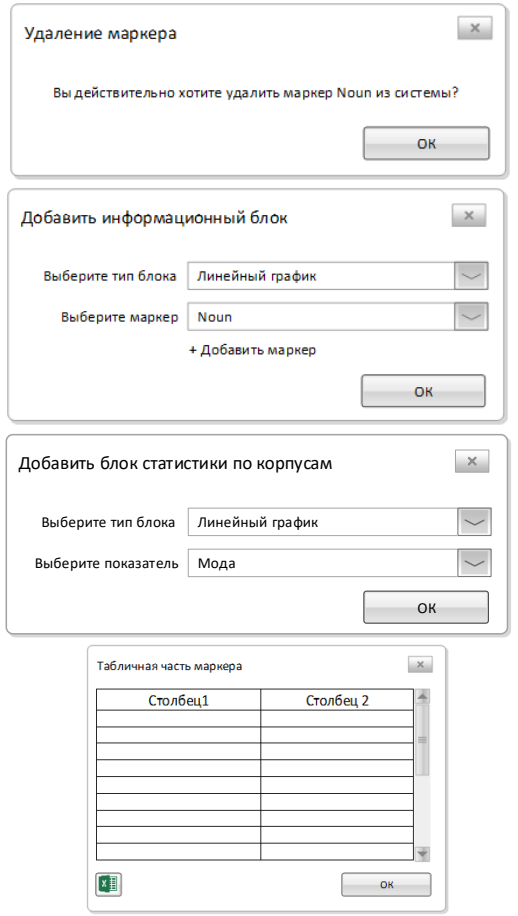
Диаграмма прецедентов





Маркер	Competent					Learners				
	BI_PE	Law	Politology	History	Economics	BI_PE	Law	Politology	History	Economics
Abstract semantic verb	0,04136	0,059857	0,06166	0,05712	0,0348091	0,07568	0,06788	0,08785	0,0849	0,189773
Archaisms	0,0000482	0,000298	0,0000781	0,00013	0,000127	0,00026	0,000059	0,0001	0,00014	0,000364
Complex conjunctions	0,00453	0,009333	0,01519	0,013	0,004345	0,00568	0,00659	0,0076	0,00905	0,016273
Desemantisized verbs	0,0281	0,046833	0,04494	0,04467	0,02226	0,05042	0,04276	0,0631	0,06157	0,123727
Future	0,00136	0,003536	0,00281	0,00157	0,001245	0,01937	0,008	0,01505	0,00729	0,035227
IPronoun	0,00035	0,006929	0,00294	0,00573	0,000536	0,00137	0,00165	0,00315	0,00638	0,013636
Intensifying adverbs	0,0017	0,002	0,00241	0,00163	0,001045	0,00137	0,001	0,00175	0,0021	0,004545
Logic connectors	0,00449	0,007369	0,0067	0,00455	0,003	0,00537	0,00641	0,009	0,00833	0,017273
Noun	0,483349	0,578214	0,76631	0,68272	0,415545	0,765	0,76147	0,9343	0,95662	0,872658
OrSuffix	0,00422	0,004131	0,0055	0,00433	0,004791	0,00637	0,1024	0,0114	0,01014	0,016409
PassiveVoice	0,02434	0,019571	0,02156	0,02467	0,011809	0,05011	0,02835	0,03965	0,03614	0,086864
Past	0,02002	0,027202	0,04452	0,0981	0,011682	0,01811	0,01647	0,02755	0,05633	0,054182
PersonalPronoun	0,01858	0,043012	0,03313	0,06252	0,0131636	0,02253	0,03394	0,03935	0,04867	0,059864
Prepositive attributes	0,22371	0,243179	0,33511	0,28315	0,199064	0,33742	0,30912	0,374	0,35971	0,719636
Postpositive attributes	0,09554	0,131167	0,15916	0,1343	0,078891	0,14274	0,182	0,20845	0,20457	0,347591
Present	0,07999	0,10356	0,09561	0,05507	0,0555	0,12426	0,11359	0,14135	0,1261	0,277727
ThatThose	0,00022	0,000524	0,00034	0,00063	0,000209	0,00005	0	0,0001	0,00014	0
WePronoun	0,01086	0,00956	0,01063	0,00555	0,007818	0,00579	0,1735	0,01555	0,01057	0,010545
YouHeShe	0,00011	0,000238	0,00028	0,00032	0,000109	0,00074	0,001	0,0012	0,00129	0,001955





Проектирование интерфейса подсистемы

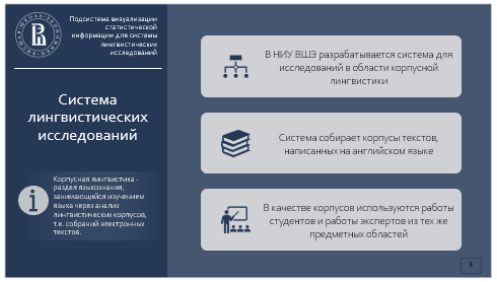
Подсистема визуализации статистической информации для системы лингвистических исследований

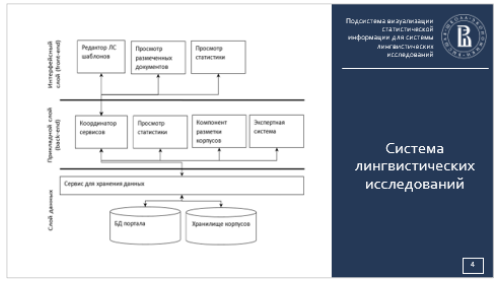
Подсистема визуализации статистической информации
для системы лингвистических исследований

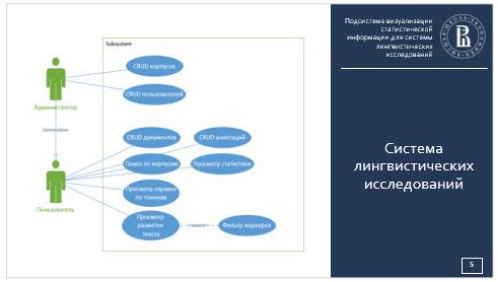
СПАСИБО ЗА ВНИМАНИЕ!

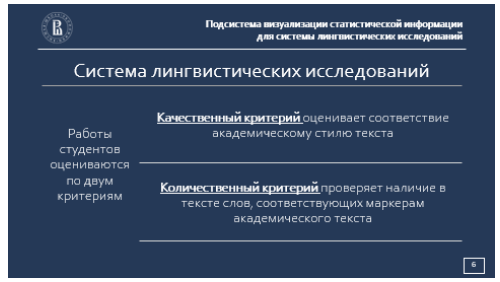
ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Содержание

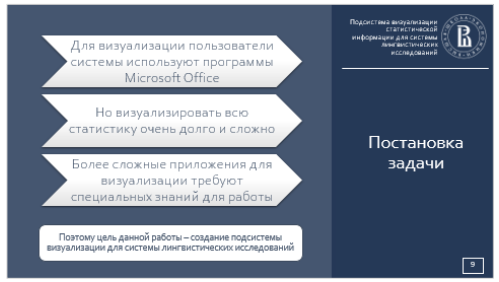

Система лингвистических исследований
В НИУ ВШЭ разрабатывается система для исследований в области корпусной лингвистики
Система собирает корпусы текстов, написанных на английском языке
В качестве корпусов используются работы студентов и работы экспертов из тех же предметных областей

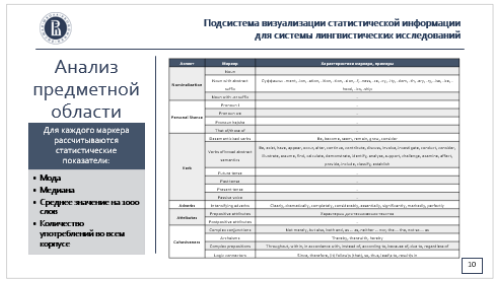

Редактор ЛС шаблонов
Просмотр размеченных документов
Просмотр статистики
Координатор сервисов
Просмотр статистики
Компонент разметки корпусов
Экспертная система
Сервис для хранения данных
БД портала
Хранилище корпусов
Система лингвистических исследований

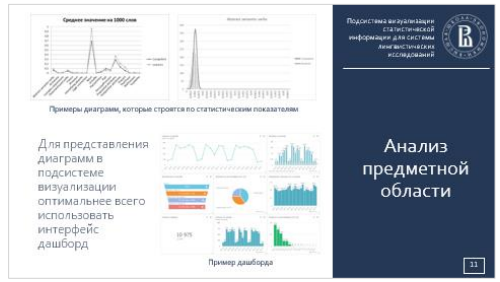

Subsystem
CRUD корпусов
CRUD пользователей
CRUD документов
CRUD аннотаций
Поиск по корпусам
Просмотр статистики
Просмотр справки по тэгам
Просмотр разметки текста
Фильтр маркеров
Система лингвистических исследований

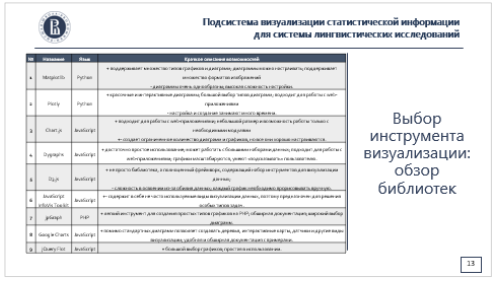

Подсистема визуализации статистической информации для системы лингвистических исследований
Система лингвистических исследований
Работы студентов оцениваются по двум критериям
Качественный критерий оценивает соответствие академическому стилю текста
Количественный критерий проверяет наличие в тексте слов, соответствующих маркерам академического текста


В процессе работы система собирает большое количество статистики по корпусам
Отчеты по корпусам получаются громоздкими и сложными для восприятия и сравнения
Для упрощения восприятия статистики необходимо визуализировать информацию – представить ее в виде графиков и диаграмм
Постановка задачи


Для визуализации пользователи системы используют программы Microsoft Office
Но визуализировать всю статистику очень долго и сложно
Более сложные приложения для визуализации требуют специальных знаний для работы
Поэтому цель данной работы – создание подсистемы визуализации для системы лингвистических исследований
Постановка задачи

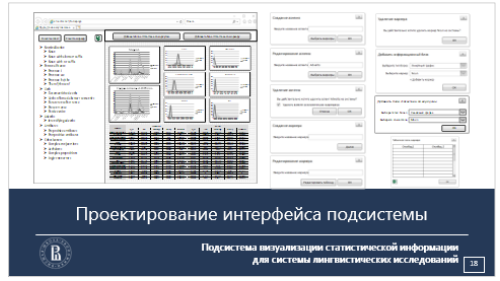

Анализ предметной области
Для каждого маркера рассчитываются статистические показатели:
Мода
Медиана
Среднее значение на 1000 слов
Количество употреблений по всем корпусе


Примеры диаграмм, которые строятся по статистическим показателям
Для представления диаграмм в подсистеме визуализации оптимальнее всего использовать интерфейс дашборд
Пример дашборда
Анализ предметной области


Выбор инструмента визуализации: обзор библиотек


Выбор инструмента визуализации: оценка библиотек
от 0 – признак отсутствует или выражен очень слабо до 10 – признак сильно выражен


Диаграмма прецедентов


Проектирование интерфейса подсистемы