Варианты задания по разработке программ АОТ

Вариант 1. "Токены"

Написать программу для подсчета статистики употребления в русскоязычном тексте токенов нескольких видов (например, имен-фамилий, названий организаций, дат и др.). Программа должна:

- осуществлять графематический анализ текста;
- составлять частотный список токенов каждого вида;
- подсчитывать абсолютную и относительную частоту токенов каждого из рассматриваемых видов;
- выводить подсчитанную статистику в удобной, обозримой форме: количество и процент токенов каждого вида, а также примеры токенов наиболее частых видов, наиболее частотные токены и др.

Отчет: указание видов выделяемых токенов и обработанных текстов, программа с комментариями, подсчитанная статистика и комментарии к способу ее подсчета, выводы.
Прикладные задачи: извлечение информации из текстов ("именованные сущности").

Вариант 2. "Буквенная статистика"

Составить программу подсчета относительной частоты используемых в тексте букв/символов (униграмм), а также символьных биграмм и триграмм. Язык текста не обязательно должен быть русским, но тексты должны быть достаточно объемными.

Программа должна:

- строить частотный словарь униграмм/биграмм/триграмм;
- выводить по запросу 3-10 наиболее частотных и наименее частотных униграмм/биграмм/триграмм;
- определять пары униграмм/биграмм/триграмм с близкой (почти равной) частотой.

Сравнить полученную статистику униграмм с известной эталонной статистикой (один из способов: сравнение графиков распределений частот).
Отчет: программа с комментариями, указание обработанных текстов, подсчитанная статистика, комментарии к способу ее подсчета, выводы по сравнению.
Прикладные задачи: распознавание языка текста, определение кодировки текста, дешифровка текста, оценка схожести текстов.

Вариант 3. "Морфология"

Составить программу, выполняющую морфологический анализ словоформ заданного русскоязычного текста и вычисление 10-12 статистических характеристик разного типа:

- морфологические: процент разных частей речи, наиболее частотные падежи у существительных и прилагательных, относительная частота падежей, наиболее частотные морфологические формы глаголов (время /лицо/ число) и т.п.
- общестатистические: общее число словоупотреблений, число различных (уникальных) словоформ/лемм, длины словоформ и их средние значения и т.п.

Следует выводить подсчитанные характеристики в удобной, обозримой форме. Предлагается рассмотреть тексты 2 разных авторов/стилей/жанров и сравнить полученные результаты.
Отчет: программа с комментариями, указание обработанных текстов, подсчитанная статистика, комментарии к способу ее подсчета, выводы по сравнению.
Прикладные задачи: определение и оценка стиля/жанра/схожести текста.

Вариант 4. "Лексикон"

Написать программу, определяющую характеристики лексикона (словарного состава) русскоязычного текста. Программа должна выполнять морфологический анализ словоформ текста и на этой основе:

- вычислять несколько общестатистических характеристик: общее число словоупотреблений, самые частотные слова (кроме стоп-слов) и их относительная частота, длины словоформ и их средние значения и т.п.;

- оценивать особенности лексикона (7-12 характеристик): количество уникальных слов (лемм), число уникальных лемм разных частей речи (существительных, прилагательных, глаголов, наречий и др.), самые частотные слова основных частей речи, коэффициент лексического богатства текста (= отношение числа различных лемм к общему числу словоупотреблений) и т.п.;
- выводить подсчитанные характеристики в удобной, обозримой форме.

Предлагается рассмотреть тексты 2 разных авторов/стилей/жанров и сравнить полученные результаты.
Отчет: программа с комментариями, указание обработанных текстов, подсчитанная статистика, комментарии к способу ее подсчета, выводы по сравнению.
Прикладные задачи: определение стиля/жанра/автора/схожести текста, составление частотных словарей, оценка развития лексики языка.

Вариант 5. "Сравнение языков"
Составить программу, сравнивающую словарный состав (лексикон) "параллельных" текстов на двух разных языках (например, русском и английском) на основе текстовой статистики. Для сравнения необходимо взять два достаточно объемных текста, один из которых является переводным эквивалентом второго (например, художественное произведение или учебник). Необходимо выполнить морфологический анализ обоих текстов, после чего программа должна:

- составлять частотный словарь языка каждого текста, оценивать его объем (т.е. количество уникальных слов/лемм), определять самые частотные слова;
- оценивать особенности словарей: число уникальных лемм разных частей речи (существительных, прилагательных, глаголов, наречий и др.), самые частотные слова основных частей речи, коэффициент лексического богатства текста (=отношение числа различных лемм к общему числу словоупотреблений) и т.п.;
- выводить подсчитанные характеристики в удобной, форме, позволяющей произвести сравнение вычисленных статистических характеристик (например, в виде таблиц или графиков сравнения).

Отчет: программа с комментариями, характеристика обработанных текстов, подсчитанная статистика, комментарии к способу ее подсчета, выводы по сделанным оценкам.
Прикладные задачи: составление частотных словарей, выявление языковых отличий и сложностей текстов.

Вариант 6. "Тональность"
Составить программу, определяющую для заданного русскоязычного текста его тональность, т.е. степень использования в нем оценочных (тональных) слов, например: "замечательный", плохой", "кринжовый" и т.п.. Тональность обычно оценивается как соотношение слов с положительной и отрицательной оценкой. Для оценки тональности текста необходимо провести морфологический анализ словоформ текста, а также вычисление нескольких его общестатистических характеристик: общее число словоупотреблений, число различных (уникальных) слов (лемм) и т.п.
Подсчитанные характеристики и оценки следует выводить в удобной, обозримой форме.
Словарь тональных слов высылается по запросу (или берется из Интернета или даже составляется самостоятельно). Описание известных тональных словарей см. на https://docs.google.com/spreadsheets/d/1wAEHOJM5LhuEpm4d6ehIukI-ecNfmPahbs1J6FxEL4c/edit?usp=sharing
Предлагается рассмотреть 2-5 текстов разных авторов/стилей/жанров и сравнить полученные оценки.
Отчет: указание исследованных текстов и примененного словаря оценочных слов, программа с комментариями, подсчитанная статистика, пояснения к способу ее подсчета, выводы по сравнению.
Прикладные задачи: выявление особенностей стиля/жанра/автора текста, оценка тональности текста.

Вариант 7. "Окрашенность"

Составить программу, оценивающую употребление в заданном русскоязычном тексте слов-названий различных цветов (белый, синий, бордовый, пурпурный, болотный и т.п.). Программа должна:

- строить словарь разных цветов, встретившихся в тексте, и встречаются ли в нем сложные цвета (светло-голубой, лимонно-желтый, лазурно-синий и др.);
- вычислять абсолютную и относительную частоту употребления каждого слова-названия цвета, наиболее часто встречающийся цвет и т.п.;
- выводить подсчитанную статистику в удобной, обозримой форме.

Для этого необходимо выполнить морфологический анализ словоформ текста, а также вычислить несколько общестатистических характеристик текста (общее число словоупотреблений, число различных, т.е. уникальных лемм и т.п.).

Словарь различных цветов можно взять из Интернета, например: https://textometr.ru/colors или же составить его самостоятельно. Предлагается рассмотреть 2-5 художественных текстов разных авторов и сравнить полученные результаты.

Отчет: указание исследованных текстов, описание примененного словаря оценочных слов, программа с комментариями, подсчитанная статистика, пояснения к способу ее подсчета, выводы по сравнению.

Прикладные задачи: выявление особенностей стиля/ жанра/автора текста.

Вариант 8. "Омонимия"

Написать программу подсчета статистических характеристик русскоязычного текста, иллюстрирующих явление омонимии. Программа должна:

- осуществлять морфологический анализ словоформ текста, в режиме без автоматического разрешения омонимии;
- вычислять:
  − общее количество словоупотреблений, число незнакомых (морфоанализатору) слов;
  − относительную частоту всех омонимичных словоформ в тексте (с учетом/без учета неизменяемых слов), среднее число омонимов у словоформ;
  − относительную частоту по разным видам омонимии: лексико-морфологической и частеречной, собственно морфологической омонимии, омонимии только по лемме;
- определять наиболее частотные омонимы, словоформы с наибольшим числом омонимов, примеры омонимов разного вида;
- выводить подсчитанную статистику в удобной, обозримой форме.

Лексико-морфологическая омонимия − совпадение некоторых грамматических форм разных слов (лексем), частеречная − различие только в части речи (а все формы одинаковые), собственно морфологическая омонимия − это совпадение некоторых грамматических форм для одного и того же слова.

Предлагается применить программу к объемному тексту (например, художественному).

Отчет: программа с комментариями, указание обработанного текста, подсчитанная статистика по видам омонимии, комментарии к способу ее подсчета, выводы.

Прикладные задачи: разрешение (снятие) морфологической и лексической омонимии (например, контекстное).

Вариант 9. "Законы текста"

Написать программу, позволяющую проверять для заданного текста выполнение закона Ципфа-Мальдеброта и/или Хипса. После выполнения морфологического анализа словоформ текста программа должна:

- подсчитывать частоты и ранги различных словоформ/лемм;
- выводить по запросу 5-20 самых частотных словоформ/лемм;
- строить график зависимости относительной частоты от ранга словоформы/леммы и/или график закона Хипса.

Для проверки закона Хипса следует рассмотреть в обрабатываемом тексте текстовые фрагменты последовательно увеличивающегося размера (в токенах/словоупотреблениях) и построить график зависимости числа уникальных лемм фрагмента от его размера.

Отчет: программа с комментариями, указание обработанных текстов, подсчитанная статистика, комментарии к способу ее подсчета, графики и выводы.
Прикладные задачи: оценка естественности текста, определение стиля/жанра текста.

Вариант 10. "Удобочитаемость"
Составить программу, вычисляющую оценку удобочитаемости русскоязычного текста по 2-3 разным формулам (индексам), и провести сравнение таких оценок на 5-7 текстах среднего размера, разных стилей и взятых из разных источников.
Выбранные для реализации формулы должны включать индекс Флэша (адаптированный для русского языка), среди других индексов можно взять также адаптированные для русского: ARI, FKGL, SMOG, CLI (см. их описание на https://www.dialog-21.ru/media/5933/veselovasplusetal066.pdf, стр. 3).
Программа должна вычислять нужные статистические характеристики текста (число букв и предложений, среднюю длину слов и предложений, среднее число слогов в слове и др.) и на этой основе оценивать удобочитаемость по выбранным индексам. Сравнение вычисленных индексов для одного и того текста следует дополнить их сравнением с экспертной оценкой человека (индивидуальным ощущением сложности).
Отчет: краткое описание выбранных индексов удобочитаемости, программа с комментариями, оцениваемые тексты, вычисленные оценки, выводы по проведенному сравнению.
Прикладные задачи: оценка сложности/удобочитаемости текста, упрощение текста.

Вариант 11. "Любимые слова"
Написать программу, определяющую любимые слова у выбранного автора русскоязычной литературы (художественной или научно-популярной, например, у Ф. Достоевского или А. Маркова). Предполагается рассмотреть достаточно объемные произведения выбранного автора, или даже несколько текстов. Любимым словом считается часто употребляемое в текстах слово, для которого выполнены следующие условия:

- слово знаменательной части речи (существительное, прилагательное, глагол, наречие), для существительных оно не должно быть именем собственным, географическим названием или именем другого типа;
- слово, которое встречается в обрабатываемых текстах не менее 1 раза на 1 тыс. слов (или 10 тыс. слов − этот порог зависит от объема обрабатываемых текстов);
- слово не должно быть очень редким: частота его употребления по данным НКРЯ или словаря Частотный словарь не должно быть меньше 1 раза на миллион слов (*ipm* − количество вхождений слова на один миллион слов текстового корпуса);
- слово присутствует в половине произведений автора (если рассматривается более одного текста/произведения).

Программа должна выполнять лемматизацию текста, подсчет статистики (включая общее число словоупотреблений в тексте/текстах), составление частотного словаря, а также определение любимых слов: вне зависимости от части речи, а также по частям речи: любимого наречия, глагола, существительного, прилагательного − по 3-5 наиболее частотных любимых слов каждой части речи. Если наиболее частотные слова не удовлетворяют указанным выше условиям, программа должна вывести соответствующее сообщение.
Отчет: программа с комментариями, характеристика обработанных текстов, описание способа подсчета статистики и определения любимых слов, выводы по результатам.
Прикладные задачи: выявление характерных черт авторского лексикона, определение стиля автора.

Вариант 12. "Засоренность речи"
Составить программу, вычисляющую для нескольких русскоязычных интернет-текстов степень их "засоренности" жаргонными/бранными/слэнговыми словами, англицизмами и др. (*дропнуть, ивент, аттач, кринжовый, фигово* и др.). Степень засоренности следует

оценивать как долю (процент) подобных слов в общем числе знаменательных слов теста (существительных, прилагательных, глаголов, наречий).

Для анализа предлагается выбрать тексты из сети Интернет (блоги, тексты Хабра и т.п.) − 5-7 текстов из разных источников и сравнить полученные оценки.

Для анализа и оценки текстов следует провести их лемматизацию, а также вычисление нескольких общестатистических характеристик:

- общее число словоупотреблений в тексте/текстах;
- количество слов разных частей речи,
- число различных (уникальных) слов (лемм) и т.п.

При вычислении нужных оценок необходимо использовать словари, например: Рабочий слэнг, и можно самостоятельно сформировать более полный словарь из нескольких источников. Также целесообразно по результатам работы морфоанализатора сформировать список несловарных слов (слов, незнакомых морфологическому анализатору), т.к. среди них могут встретиться новые жаргонные и слэнговые слова.

Отчет: описание обработанных текстов и использованных словарей, программа с комментариями, результирующие оценки засоренности текстов, пояснения к способу их подсчета, выводы по результатам.

Прикладные задачи: выявление характерных особенностей текстов, оценка соответствия лексики текста литературной норме.