Домашнее задание №3

Дудырев Егор

10.03.2020

1 Задание

Задание E
Составить (с использованием любого модуля морфоанализа) программу, выполняющую извлечение словосочетаний определенного вида из заданного русскоязычного текста. Выделение словосочетаний может базироваться на локальных высоковероятных синтаксических связях (см. слайд 50 Лекции № 7). Программа выводит все словосочетания заданного вида/ов, встречающиеся в обрабатываемом тексте. Рассмотреть несколько (2-5) грамматических образцов словосочетаний, например:

1. именные словосочетания (NP), в которые входят грамматически согласованные прилагательные и существительное (и, возможно, порядковые числительные или наречие, например: первый детальный написанный план или более сложный образец: первый детально написанный план);

2. именные словосочетания, включающие несколько существительных и прилагательные (зависимые существительные входят в родительном падеже, прилагательные согласованы с соответствующими существительными, например: легкий отблеск далекого заката);

3. предложные словосочетания, в которые входят именные словосочетания одного из грамматических образцов (например: в тихом летнем парке);

4. глагольные группы, состоящие из глагола в личной форме и зависимой именной группы (например: увидел большой стеклянный шар).

Протестировать программу на нескольких небольших текстах разных жанров.

Отчет: Описание грамматических образцов извлекаемых словосочетаний и стратегии (алгоритма) их выделения; составленная и примененная программа с комментариями, результаты ее тестирования.

2 Используемые данные

Используемые тексты:

- Художественные тексты
    1. Глава 1 книги "Алиса в Стране Чудес"Л. Кэрролла
    2. Рассказ "Крыжовник"А.П. Чехова
    3. Глава 1 книги "Трудно быть богом"братьев Стругацких
- Научно-деловые тексты
    1. Главы 1, 2 книги "Пособие по написанию разного рода деловых текстов"Г.П. Несговорова
    2. Статья "Алгоритм помог расслышать гравитационные волны от землетрясений без суперЭВМ"сайта N+1
    3. Статья "Антиматерия не отличилась от материи взаимодействием с квантовыми флуктуациями"сайта N+1
    4. Статья "Универсальность математики"Ф.Х. Айматовой

3 Используемое ПО

Язык программирования: Python 3.7 (с дистрибутивом Anaconda).
В качестве лексического анализатора использовалась библиотека pymorphy2 0.8
Используемые библиотеки:

- numpy 1.17.2
- pandas 0.25.1
- matplotlib 3.1.1
- seaborn 0.9.0

- sklearn 0.21.3
- pymystem3 0.2.0

4 Описание грамматических образцов извлекаемых словосочетаний и алгоритмы их выделения

В качестве грамматических образцов рассмотрим примеры из описания задания. Для урощения алгоритма поиска весь текст разбивается на предложения. Это можно сделать, т.к. все определяемые грамматические образцы могут находиться только в пределах одного предложения.

4.1 Именные словосочетания с одним существительным

Описание
Именные словосочетания (NP), в которые входят грамматически согласованные прилагательные и существительное (и, возможно, порядковые числительные или наречие, например: первый детальный написанный план или более сложный образец: первый детально написанный план)
Алгоритм поиска

1. Найти существительные в предложении
2. Для каждого существительного:
   (a) Посмотреть на слово слева от текущего
   (b) Если слово - прилагательное, причастие, порядковое числительное или наречие и согласовано с существительным, запомнить его и перейти к п.2(a)
   (c) Иначе закончить обработку существительного с запомненными для него прилагательными и причастиями

4.2 Именные словосочетания с зависимыми существительными

Описание
Именные словосочетания, включающие несколько существительных и

прилагательные (зависимые существительные входят в родительном падеже, прилагательные согласованы с соответствующими существительными, например: легкий отблеск далекого заката);
Алгоритм поиска

1. Находим все именные словосочетания с одним существительным
2. Смотрим на все варианты перестановок (powerset) множества словосочетаний из п.1
   (a) Если конкатенация текущего варианта перестановки словосочетаний отсутствует в исходном предложении - переходим к следующему варианту.
   (b) Если конкатенация текущего варианта перестановок словосочетаний присутствует в исходном предложении:
      i. Проверяем, что одно из существительных в текущем варианте перестановок находится не в родительном падеже
      ii. Проверяем, что все остальные существительные в текущем варианте перестановок находятся в родительном падеже
      iii. Если оба условия соблюдены - то текущая перестановка словосочетаний является именным словосочетанием с зависимыми существительными
3. Среди найденных именных словосочетаний с зависимыми существительными оставляем только наиболе общие, т.е. словосочетания не являющиеся частью других словосочетаний

4.3 Предложные словосочетания

Описание
Предложные словосочетания, в которые входят именные словосочетания одного из грамматических образцов (например: в тихом летнем парке).
Алгоритм поиска

1. Находим все именные словосочетания с одним существительным и с зависимыми существительными.
2. Среди найденных именных словосочетаний оставляем только наиболе общие, т.е. словосочетания не являющиеся частью других словосочетаний
3. Проходим по каждому именному словосочетанию:

(a) Если слово слева от словосочетания - предлог, то словосочетание - предложное.

4.4 Глагольные группы

Описание
Глагольные группы, состоящие из глагола в личной форме и зависимой именной группы (например: увидел большой стеклянный шар).
Алгоритм поиска

1. Находим все именные словосочетания с одним существительным и с зависимыми существительными.
2. Среди найденных именных словосочетаний с зависимыми существительными оставляем только наиболе общие, т.е. словосочетания не являющиеся частью других словосочетаний
3. Среди найденных именных словосочетаний оставляем только наиболе общие, т.е. словосочетания не являющиеся частью других словосочетаний
4. Оставляем словосочетания в которых содержатся существительные в винительном падеже - к ним должен обращаться глагол в личной форме других словосочетаний
5. Проходим по каждому оставшемуся словосочетанию:
   (a) Если слово слева от словосочетания - глагол и стоит в личной форме, то словосочетание является глагольной группой.

5 Результаты тестирования

5.1 Алиса в стране чудес, Глава 1

Именные словосочетания с одним существительным
Количество словосочетаний: 72
Средняя длина словосочетаний: 2.17
Примеры:

- "Вдруг мимо пробежал кролик с красными глазами" $\Longrightarrow$ "красными глазами"
- "однако в тот миг все казалось ей вполне естественным" $\Longrightarrow$ "тот миг"

- "когда Кролик вдруг вынул часы из жилетного кармана и" $\Longrightarrow$ "жилетного кармана"
- "да еще с жилетным карманом в придачу" $\Longrightarrow$ "жилетным карманом"
- "словно в глубокий колодец" $\Longrightarrow$ "глубокий колодец"

Именные словосочетания с зависимыми существительными
Количество словосочетаний: 1
Средняя длина словосочетаний: 4.00
Примеры:

- "Однако на этом пузырьке никаких пометок не было" $\Longrightarrow$ "этом пузырьке никаких пометок"

Предложные словосочетания
Количество словосочетаний: 29
Средняя длина словосочетаний: 3.23
Примеры:

- "Вдруг мимо пробежал кролик с красными глазами" $\Longrightarrow$ "с красными глазами"
- "однако в тот миг все казалось ей вполне естественным" $\Longrightarrow$ "в тот миг"
- "когда Кролик вдруг вынул часы из жилетного кармана и" $\Longrightarrow$ "из жилетного кармана"
- "да еще с жилетным карманом в придачу" $\Longrightarrow$ "с жилетным карма

Глагольные группы
Количество словосочетаний: 9
Средняя длина словосочетаний: 3.00
Примеры:

- "хоть сейчас был не самый подходящий момент демонстрировать свои познания" $\Longrightarrow$ "демонстрировать свои познания"
- "но ей очень нравились эти слова" $\Longrightarrow$ "нравились эти слова"
- "Тут раздался страшный треск" $\Longrightarrow$ "раздался страшный треск"
- "Перед ней тянулся другой коридор" $\Longrightarrow$ "тянулся другой коридор"
- "Вдруг она увидела стеклянный столик на трех ножках" $\Longrightarrow$ "увидела стеклянный столик"

5.2 Крыжовник

Именные словосочетания с одним существительным
Количество словосочетаний: 193
Средняя длина словосочетаний: 2.12
Примеры:

- "Еще с раннего утра всё небо обложили дождевые тучи" $\Longrightarrow$ "раннего утра "всё небо "дождевые тучи"
- "как бывает в серые пасмурные дни" $\Longrightarrow$ "серые пасмурные дни"
- "Ветеринарный врач Иван Иваныч и учитель гимназии Буркин уже утомились идти" $\Longrightarrow$ "Ветеринарный врач"
- "Далеко впереди еле были видны ветряные мельницы села Мироносицкого" $\Longrightarrow$ "ветряные мельницы"
- "зеленые ивы" $\Longrightarrow$ "зеленые ивы"

Именные словосочетания с зависимыми существительными
Количество словосочетаний: 1
Средняя длина словосочетаний: 4.00
Примеры:

- "Еще с раннего утра всё небо обложили дождевые тучи" $\Longrightarrow$ "раннего утра всё небо"

Предложные словосочетания
Количество словосочетаний: 76
Средняя длина словосочетаний: 3.05
Примеры:

- "Еще с раннего утра всё небо обложили дождевые тучи" $\Longrightarrow$ "с раннего утра всё небо"
- "как бывает в серые пасмурные дни" $\Longrightarrow$ "в серые пасмурные дни"
- "который издали похож на ползущую гусеницу" $\Longrightarrow$ "на ползущую гусеницу"
- "а в ясную погоду оттуда бывает виден даже город" $\Longrightarrow$ "в ясную погоду"
- "в тихую погоду" $\Longrightarrow$ "в тихую погоду"

Глагольные группы
Количество солосочетаний: 20
Средняя длина словосочетаний: 3.30
Примеры:

- "Еще с раннего утра всё небо обложили дождевые тучи" $\Longrightarrow$ "обложили дождевые тучи"
- "И минут через пять лил уже сильный дождь" $\Longrightarrow$ "лил уже сильный дождь"
- "Иван Иваныч и Буркин испытывали уже чувство мокроты" $\Longrightarrow$ "испытывали уже чувство"
- "Он сел на ступеньке и намылил свои длинные волосы и шею" $\Longrightarrow$ "намылил свои длинные волосы"
- "и на волнах качались белые лилии" $\Longrightarrow$ "качались белые лилии"

5.3 Трудно быть богом, Глава 1

Именные словосочетания с одним существительным Количество солосочетаний: 397
Средняя длина словосочетаний: 2.18
Примеры:

1. "седьмую по счету и последнюю на этой дороге" $\Longrightarrow$ "этой дороге"
2. "Хваленый хамахарский жеребец" $\Longrightarrow$ "Хваленый хамахарский жеребец"
3. "взятый у дона Тамэо за карточный долг" $\Longrightarrow$ "карточный долг"
4. "оказался сущим барахлом" $\Longrightarrow$ "сущим барахлом"
5. "вихляющейся рысью" $\Longrightarrow$ "вихляющейся рысью"

Именные словосочетания с зависимыми существительными
Количество солосочетаний: 14
Средняя длина словосочетаний: 4.29
Примеры:

1. "в лучших традициях серых казарм" $\Longrightarrow$ "лучших традициях серых казарм"

2. "У коновязи перед корчмой топтались оседланные кони серого патруля" $\Longrightarrow$ "оседланные кони серого патруля"

3. "Благородному дону счастливого пути" $\Longrightarrow$ "Благородному дону счастливого пути"

4. "что триста лет назад железные роты имперского маршала Тоца" $\Longrightarrow$ "назад железные роты имперского маршала"

5. "а разухабистые егеря барона Пампы жарили на редких полянах ворованных быков" $\Longrightarrow$ "редких полянах ворованных быков"

Предложные словосочетания
Количество словосочетаний: 118
Средняя длина словосочетаний: 3.28
Примеры:

1. "седьмую по счету и последнюю на этой дороге-> "на этой дороге"

2. "взятый у дона Тамэо за карточный долг-> "за карточный долг"

3. "В мутном небе дрожали редкие тусклые звезды-> "В мутном небе"

4. "как всегда осенью в этой приморской стране с душными-> "в этой приморской стране"

5. "одна из бесчисленных однообразных Мертвожорок-> "из бесчисленных однообразных Мертвожорок"

Глагольные группы
Количество солосочетаний: 49
Средняя длина словосочетаний: 3.22
Примеры:

1. "В мутном небе дрожали редкие тусклые звезды" $\Longrightarrow$ "дрожали редкие тусклые звезды"

2. "По сторонам тянулись распаханные поля" $\Longrightarrow$ "тянулись распаханные поля"

3. "Далеко слева вспыхивало и гасло угрюмое зарево" $\Longrightarrow$ "гасло угрюмое зарево"

4. "Почему бы горожанину Киуну не найти бескорыстную защиту у глупого и спесивого аристократа" $\Longrightarrow$ "найти бескорыстную защиту"

5. "Я знавал одного Киуна" $\Longrightarrow$ "знавал одного Киуна"

5.4 Пособие по написанию разного рода деловых текстов

Именные словосочетания с одним существительным
Количество солосочетаний: 346
Средняя длина словосочетаний: 2.16
Примеры:

1. "Научным сотрудникам" $\Longrightarrow$ "Научным сотрудникам"
2. "инженерам и людям других творческих специальностей в своей профессиональной деятельности не обойтись без оформления ряда документов" $\Longrightarrow$ "других творческих специальностей "своей профессиональной деятельности"
3. "разного рода описания" $\Longrightarrow$ "разного рода"
4. "деловые письма" $\Longrightarrow$ "деловые письма"
5. "При этом надо учитывать особенности деловой письменной речи" $\Longrightarrow$ "деловой письменной речи"

Именные словосочетания с зависимыми существительными
Количество солосочетаний: 31
Средняя длина словосочетаний: 4.35
Примеры:

1. "и следовало бы изучить особенности деловой разновидности родного русского языка" $\Longrightarrow$ "деловой разновидности родного русского языка"
2. "В предлагаемом пособии за основу взят не традиционный курс русского языка" $\Longrightarrow$ "традиционный курс русского языка"
3. "изучает письменную форму официальноделового общения с целью ее унификации и нормализации" $\Longrightarrow$ "письменную форму официальноделового общения"
4. "их разработке и применении или о моделировании с их помощь азличных реальных систем" $\Longrightarrow$ "их помощь азличных реальных систем"
5. "Можно отметить следующие особенности информационной стилистики" $\Longrightarrow$ "следующие особенности информационной стилистики"

Предложные словосочетания
Количество солосочетаний: 99
Средняя длина словосочетаний: 3.35
Примеры:

1. "инженерам и людям других творческих специальностей в своей профессиональной деятельности не обойтись без оформления ряда документов" $\Longrightarrow$ "в своей профессиональной деятельности"

2. "Если для художественной прозы характерна многозначность" $\Longrightarrow$ "для художественной прозы"

3. "При этом в деловых текстах используются развернутые синтаксические конструкции и точное употребление слов" $\Longrightarrow$ "в деловых текстах"

4. "чем отличаются деловые тексты от разговорной речи или художественной прозы" $\Longrightarrow$ "от разговорной речи"

5. "какие конструкции или выражения неприемлемы в деловых текстах" $\Longrightarrow$ "в деловых текстах"

Глагольные группы
Количество солосочетаний: 30
Средняя длина словосочетаний: 3.58
Примеры:

1. "При этом в деловых текстах используются развернутые синтаксические конструкции и точное употребление слов" $\Longrightarrow$ "используются развернутые синтаксические конструкции"

2. "чем отличаются деловые тексты от разговорной речи или художественной прозы" $\Longrightarrow$ "отличаются деловые тексты"

3. "Для англоговорящих людей существует специальный курс Business English" $\Longrightarrow$ "существует специальный курс"

4. "что при составлении различного рода документации используются несколько иные речевые нормы ем в литературном или бытовом языке" $\Longrightarrow$ "используются несколько иные речевые нормы"

5. "которая устанавливает речевые нормы при написании разного рода документации" $\Longrightarrow$ "устанавливает речевые нормы"

5.5 Статьи N+1

Именные словосочетания с одним существительным
Количество солосочетаний: 107
Средняя длина словосочетаний: 2.16
Примеры:

1. "Геофизики разработали простой и не требовательный к вычислительным ресурсам программный инструмент" $\Longrightarrow$ "вычислительным ресурсам "программный инструмент"

2. "который позволяет идентифицировать гравитационные волны" $\Longrightarrow$ "гравитационные волны"

3. "Поскольку гравитационные волны распространяются со скоростью света" $\Longrightarrow$ "гравитационные волны"

4. "новый метод поможет детектировать землетрясения намного раньше" $\Longrightarrow$ "новый метод"

5. "чем к сейсмостанциям придут обычные сейсмические волны" $\Longrightarrow$ "обычные сейсмические волны"

Именные словосочетания с зависимыми существительными
Количество солосочетаний: 11
Средняя длина словосочетаний: 4.18
Примеры:

1. "но для создания новых систем раннего оповещения о сейсмических толчках необходимо создать новые более чувствительные гравиметры и инфраструктуру" $\Longrightarrow$ "новых систем раннего оповещения"

2. "Физики измерили тонкое расщепление и лэмбовский сдвиг энергетических состояний атома антиводорода" $\Longrightarrow$ "лэмбовский сдвиг энергетических состояний"

3. "однако на больших масштабах нет никаких признаков сосуществования двух видов материи" $\Longrightarrow$ "больших масштабах нет никаких признаков"

4. "Факт столь значительного преобладания обычного вещества называется барионной асимметрией Вселенной" $\Longrightarrow$ "значительного преобладания обычного вещества"

5. "то есть эквивалентности физических процессов при одновременной инверсии всех зарядов" $\Longrightarrow$ "одновременной инверсии всех зарядов"

Предложные словосочетания
Количество солосочетаний: 37
Средняя длина словосочетаний: 3.35
Примеры:

1. "Геофизики разработали простой и не требовательный к вычислительным ресурсам программный инструмент" $\Longrightarrow$ "к вычислительным ресурсам"

2. "но сигнал от землетрясений не удавалось вычленить из фоновых шумов и артефактов" $\Longrightarrow$ "из фоновых шумов"

3. "Во время землетрясения в японской обсерватории Камиока работал сверхпроводящий гравиметр" $\Longrightarrow$ "в японской обсерватории"

4. "Однако в этом случае речь шла об исключительно мощном землетрясении" $\Longrightarrow$ "в этом случае "об исключительно мощном землетрясении"

5. "Решить эту задачу взялись ученые из германского Центра наук о Земле в Потсдаме и университета Пекина под руководством Себастьяна Хайманна" $\Longrightarrow$ "из германского Центра"

Глагольные группы
Количество солосочетаний: 25
Средняя длина словосочетаний: 3.42
Примеры:

1. "который позволяет идентифицировать гравитационные волны" $\Longrightarrow$ "идентифицировать гравитационные волны"

2. "чем к сейсмостанциям придут обычные сейсмические волны" $\Longrightarrow$ "придут обычные сейсмические волны"

3. "Землетрясения провоцируют быстрое перераспределение масс в недрах Земли и порождают сейсмические волны" $\Longrightarrow$ "провоцируют быстрое перераспределение "порождают сейсмические волны"

4. "то есть гравитационные волны" $\Longrightarrow$ "есть гравитационные волны"

5. "Еще в начале 2000 годов ученые пытались обнаружить эти возмущения" $\Longrightarrow$ "обнаружить эти возмущения"

5.6 Универсальность математики

Именные словосочетания с одним существительным
Количество солосочетаний: 85
Средняя длина словосочетаний: 2.20
Примеры:

1. "Основные задачи изучения математики как науку" $\Longrightarrow$ "Основные задачи"
2. "являются оперирующими чистыми абстракциями" $\Longrightarrow$ "оперирующими чистыми абстракциями"
3. "отделёнными от реального мира" $\Longrightarrow$ "реального мира"
4. "Воспитание математической культуры" $\Longrightarrow$ "математической культуры"
5. "развитие логического и алгоритмического мышления" $\Longrightarrow$ "алгоритмического мышления"

Именные словосочетания с зависимыми существительными
Количество солосочетаний: 3
Средняя длина словосочетаний: 4.33
Примеры:

1. "Но теории и методы созданные в рамках таких формализованных систем могут найти неожиданное применение в различных отраслях научного знания" $\Longrightarrow$ "различных отраслях научного знания"
2. "второй вид полной формализации" $\Longrightarrow$ "второй вид полной формализации"
3. "их существенная зависимость от случайных природных факторов обуславливают вероятностный характер многих производственных процессов" $\Longrightarrow$ "вероятностный характер многих производственных процессов"

Предложные словосочетания
Количество солосочетаний: 16
Средняя длина словосочетаний: 3.35
Примеры:

1. "отделёнными от реального мира" $\Longrightarrow$ "от реального мира"

2. "развитие способности к дальнейшему самостоятельному образованию" $\Longrightarrow$ "к дальнейшему самостоятельному образованию"

3. "Математика стала широко проникать во все сферы науки" $\Longrightarrow$ "во все сферы"

4. "А саму математику он рассматривает как науку о хитроумных операциях" $\Longrightarrow$ "о хитроумных операциях"

5. "производимых по специально разработанным правилам над специально придуманными понятиями" $\Longrightarrow$ "по специально разработанным правилам "над специально придуманными понятиями"

Глагольные группы
Количество солосочетаний: 10
Средняя длина словосочетаний: 3.40
Примеры:

1. "использовать абстрактные математические модели для изучения конкретных процессов и явлений" $\Longrightarrow$ "использовать абстрактные математические модели"

2. "из которых затем выводится основное содержание теории" $\Longrightarrow$ "выводится основное содержание"

3. "и объекты и аксиомы имеют свои аналоги в мире вещей" $\Longrightarrow$ "имеют свои аналоги"

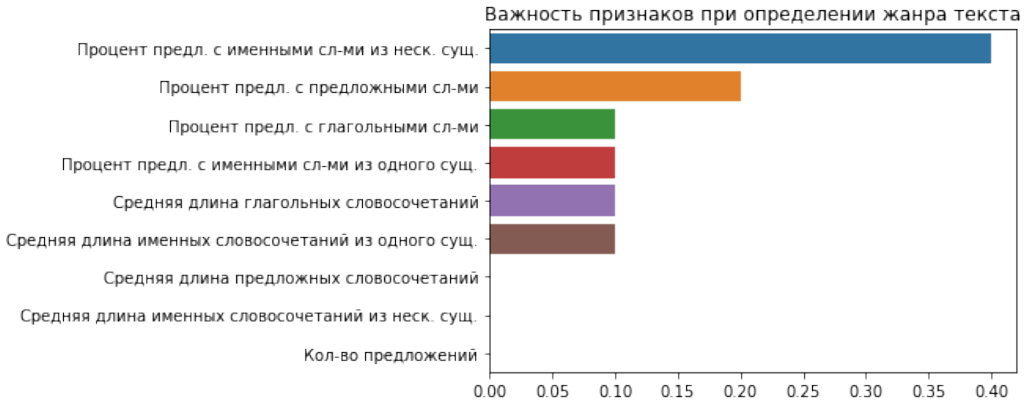
4. "что сходство начальных условий позволяет применять старую теорию для изучения новых объектов" $\Longrightarrow$ "применять старую теорию"

5. "Одно время даже пытались создать единый алгоритм для решения любых задач" $\Longrightarrow$ "создать единый алгоритм"

6 Сравнение художественных и научно-технических текстов

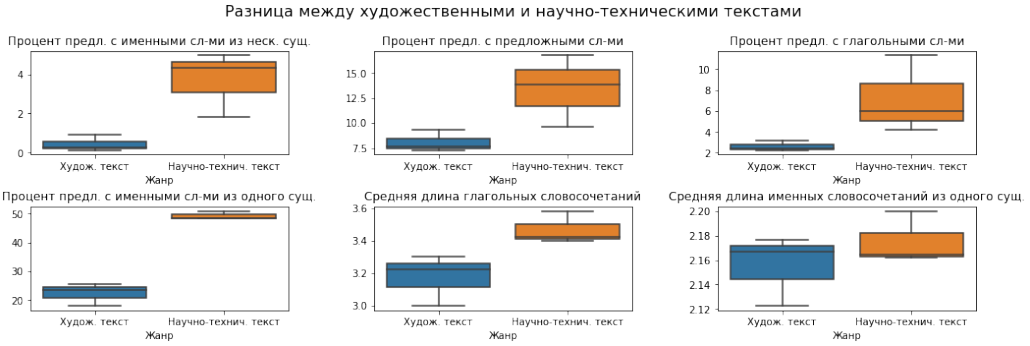
Можно ли определить жанр текста основываяь на статистике употребления в нём словосочетаний?
Обучим для этого модель Случайного леса и посмотрим на получившиеся важности признаков.

Рис. 1: Важности статистических признаков, полученные с помощью Случайного леса



Рассмотрим перечисленные важные признаки в отдельности.

Рис. 2: Распределения статистических признаков по жанрам



В научно-технических текстах содержится гораздо больше предложений со всеми видами определяемых грамматических образцов. Также длины предложных, глагольных и именных словосочетаний в научно-технических текстах значительно больше, чем в художественных. Возможно из-за длинных терминов.

7 Ссылки на код

- Исполняемый код
- Графики сравнения двух текстов