Домашнее задание №1

Дудырев Егор

24 февраля 2020 г.

1 Задание

Задание D
Выполнить лексико-статистический анализ двух текстов на русском языке, среднего размера:

1. текст художественной прозы, например, текста главы 2 из книги Л Кэрролла «Алиса в Стране Чудес» или главы из романа Л.Толстого или рассказа А.Чехова.

2. текст научно-технической или деловой прозы.

Для этого следует составить программу, которая с помощью выбранного морфопроцессора, например: Диалинг-АОТ , mystem, pymorphy или CrossMorphy (ссылки см. в варианте А), осуществляет морфологический анализ словоформ текста; вычисляет 7-15 статистических характеристик разного типа:

1. общестатистические: общее число словоупотреблений, число различных словоформ, средняя длина предложения (если процессор разбивает текст на предложения) и т.п.;

2. морфологические: абсолютная и относительная частота омонимичных словоформ, процент разных частей речи, наиболее частотные падежи у существительных и прилагательных, относительную частоту падежей, наиболее частотные морфологические формы глаголов (время/лицо/число) и т.п.;

3. лексические: количество уникальных лемм, число уникальных лемм разных частей речи (существительных, глаголов и др.), число незнакомых слов, самые частотные слова и их относительная частота, самые частотные слова основных частей речи (существительные,

прилагательные, наречия, глаголы), коэффициент лексического богатства текста (=отношение числа различных лемм к общему числу словоупотреблений) и т.п. выводит подсчитанные характеристики в удобной, обозримой форме.

Отчет: составленная программа, подсчитанная статистика (в удобной, обозримой форме, в зависимости от стилей/жанров текстов), пояснения по способу ее подсчета и выводы, программа с комментариями.

2 Используемые данные

Сравнение двух (художественного и научно-делового) текстов:

- Худ. текст: Глава 2 книги "Алиса в Стране Чудес"Л. Кэрролла
- Деловой текст: Глава 1 книги "Пособие по написанию разного рода деловых текстов"Г.П. Несговорова

Сравнение нескольких художественных и научно-деловых текстов:

- Художественные тексты
  1. Книги "Алиса в Стране Чудес"Л. Кэрролла
  2. Рассказ "Крыжовник"А.П. Чехова
  3. Рассказ "Радость"А.П. Чехова
  4. Глава 1 книги "Трудно быть богом"братьев Стругацких
  5. Глава 1 книги "Детство"Л.Н. Толстого
- Научно-деловые тексты
  1. Главы 1,2 книги "Пособие по написанию разного рода деловых текстов"Г.П. Несговорова
  2. Статья "Компьютерная лингвистика: Теория и практика"М.Р. Смагина
  3. Статья "Алгоритм помог расслышать гравитационные волны от землетрясений без суперЭВМ"сайта N+1
  4. Статья "Антиматерия не отличилась от материи взаимодействием с квантовыми флуктуациями"сайта N+1
  5. 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
- re 2.2.1
- sklearn 0.21.3

4 Сравнение научного и художественного текстов

Сильнее всего среди текстов различаются следующие характеристики слов:

- Вид глагола: В научном тексте много глаголов несовершенного вида, в художественном - совершенного
- Падежи: В научном тексте половина слов имеет родительный падеж, в художественном - именительный
- Лицо глагола: В научном тексте абсолютное большинсто глаголов - в 3ем лице. В художественном тексте анализатор pymorphy2 не может определить лица половины глаголов - вероятно они своеобразно, художественно используются.
- Части речи: В научном тексте много существительных и прилагательных, в художественном - глаголов и местоимений.
- Время глагола: В научном тексте абсолютное большинство глаголов указаны в настоящем времени. В художественном - очень много глаголов в прошедшем времени.

- Среднее количество слов в предложении научного текста примерно в 3 раза больше, чем в художественном.
- Наиболее частые леммы текстов отличаются, но являются специфичными для конкретного текста. Например в "Пособии по написанию ... текстов"очень часто встречается лемма "язык а в "Алисе в Стране Чудес имя "Алиса".

5 Сравнение нескольких текстов

Сравнить между собой два текста интересно, но ещё интереснее было бы сравнить абстрактные художественный и научно-деловой текст.
Для этого загрузим несколько различных текстов различных авторов (русские и английский художественные тексты, научные тексты по лингвистике, математике, физике), разобьём их на предложения и сгенерируем случайным образом из всех полученных предложений несколько абстрактных художественных и научно-деловых текстов. Затем обучим на них модель машинного обучения (в частности Деревья решений) и найдём главные отличия двух типов текстов.
В результате получились следующие правила, способные со стопроцентной точностью отделить художественный текст от научного:

1. Доля глаголов несов. вида > 0.67 => науч.-деловой
2. Доля глаголов сов. вида > 0.33 => художественный
3. Доля слов в род. падеже > 0.38 => науч.-деловой
4. Доля слов в 3ем лице > 0.42 => науч.-деловой
5. Доля слов с нераспознанным pymorphy2 лицом > 0.50 => художественный
6. Доля полных прилагательных > 0.15 => науч.-деловой
7. Доля существительных > 0.34 => науч.-деловой
8. Доля мест.-существительных > 0.04 => художественный
9. Доля глаголов > 0.11 => художественный
10. Доля слов в прошедшем времени > 0.47 => художественный
11. Доля слов в настоящем времени > 0.43 => науч.-деловой

12. Кол-во уникальных лемм полн. прилагательных $>$ 85.00 $=>$ науч.-деловой

6 Ссылки на код

- Исполняемый код
- Графики сравнения двух текстов