Домашнее задание №2

Алиев М.А. группа 154

Тема - извлечение устойчивых словосочетаний

Введение

Целью проекта является поиск устойчивых словосочетаний в тексте. Для реализации данного задания для анализа было выбрано произведение "Мертвые души" Н. В. Гоголя. В качестве методов рассчета "устойчивости" планируется использовать как обычный частотный метод (делим количество всех биграмм текста на количество использования целевой), так и более продвинутые метрики, такие как Dice и Mi. Каждая из них имеет как свои плюсы, так и минусы, и может лучше отрабатывать в определенных случаях.

Вся работа проводилась на языке Python3, код может быть найден в jupyter-notebook файле.

Ищем самое частое слово

Для поиска самого употребляемого слова текст для начала следует разбить на токены, а также провести их лемматизацию, чтобы не рассматривать разные формы одного и того же слова как отдельные сущности ('хорошего', 'хороший', 'хороших' и тд.). В качестве токенизатора был использован самописный инструмент, разработанный в рамках 1го домашнего задания нашего курса, для лемматизации было решено применить модуль `pymorphy2` .

После первого подсчета слов стало ясно, что с отрывом лидируют разнообразные предлоги, союзы и тд. Во избежание "замусоривания" статистики были удалены все стоп-слова, список которых был взят из модуля `NLTK` . Итоговый список приведен ниже (показаны только 10 лидирующих):

```
[('весь', 1273), ('это', 845), ('сказать', 834), ('чичиков', 780), ('который', 688),
('человек', 503), ('свой', 444), ('мочь', 411), ('дело', 406), ('говорить', 392)]
```

Было решено проводить поиск со словом ***дело***.

Разбиваем текст на биграммы

Поиск биграмм с выбранным словом проводился довольно просто: бежим по списку всех слов и если встречаем целевое, добавляем в список уже найденных 2 новых словосочетания:

`goal_word + word_N+1`

`word_N-1 + goal_word`

Считаем метрики для найденных биграмм

Начнем с самого простого - **относительной частоты употребления.**

Ниже приведены 10 лидирующих биграмм(слева направо: количество их использования, относительная частота, а также 2 самые употребляемые формы из текста)

```
'самый   дело  ' -> кол-во:  59, отн. частота: 0.00050, примеры: 'самом деле', 'самого
дела'
'дело    и     ' -> кол-во:  21, отн. частота: 0.00018, примеры: 'дело и', 'дела и'
'это     дело  ' -> кол-во:  20, отн. частота: 0.00017, примеры: 'это дело', 'этом
деле'
'дело    в     ' -> кол-во:  18, отн. частота: 0.00015, примеры: 'дело в', 'деле в'
'и       дело  ' -> кол-во:  17, отн. частота: 0.00014, примеры: 'и дело', 'и дела'
'дело    что   ' -> кол-во:  12, отн. частота: 0.00010, примеры: 'дело что', 'деле что'
'дело    быть  ' -> кол-во:  12, отн. частота: 0.00010, примеры: 'дело было', 'дело
будет'
'дело    не    ' -> кол-во:  12, отн. частота: 0.00010, примеры: 'дело не', 'деле не'
'в       дело  ' -> кол-во:  12, отн. частота: 0.00010, примеры: 'в дело', 'в деле'
'другой  дело  ' -> кол-во:  10, отн. частота: 0.00008, примеры: 'другое дело', 'других
дел'
```

Теперь воспользуемся **Dice** метрикой:

```
'самый      дело ' -> мера: 0.16643, кол-во:  59, примеры: 'самом деле', 'самого дела'
'главный    дело ' -> мера: 0.04196, кол-во:   9, примеры: 'главное дело', ' '
'это        дело ' -> мера: 0.03197, кол-во:  20, примеры: 'это дело', 'этом деле'
'другой     дело ' -> мера: 0.02538, кол-во:  10, примеры: 'другое дело', 'других дел'
'дело       идти ' -> мера: 0.02537, кол-во:   6, примеры: 'дело идет', 'дело идут'
'важный     дело ' -> мера: 0.02381, кол-во:   5, примеры: 'важным делом', 'важные
дела'
'чем        дело ' -> мера: 0.02335, кол-во:   6, примеры: 'чем дело', ' '
'бесчестный дело ' -> мера: 0.01923, кол-во:   4, примеры: 'бесчестнейшее дело',
'бесчестных дел'
'дело       но   ' -> мера: 0.01731, кол-во:   9, примеры: 'дело но', 'дел но'
'ваш        дело ' -> мера: 0.01709, кол-во:   5, примеры: 'ваше дело', 'вашему делу'
```

Заметим улучшение качества результата. Пропали словосочетания, включающие в себя союзы, местоимения и тд., потому что они часто употреблялись и без нашего ключевого слова.

Mi метрика:

```
'отлагать       дело    ' -> мера: 4.901907188094263, кол-во:   1. примеры: 'отлагая
дела'
'христолюбивый  дело    ' -> мера: 4.901907188094263, кол-во:   1. примеры:
'христолюбивое дело'
'распутывать    дело    ' -> мера: 4.901907188094263, кол-во:   1. примеры:
'распутывать дело'
'премерзкий     дело    ' -> мера: 4.901907188094263, кол-во:   1. примеры:
'премерзейшее дело'
'дело           беспримерный' -> мера: 4.901907188094263, кол-во:   1. примеры: 'деле
беспримерное'
'обделывать     дело    ' -> мера: 4.901907188094263, кол-во:   1. примеры: 'обделывать
дела'
'запутаннейший  дело    ' -> мера: 4.901907188094263, кол-во:   2. примеры:
'запутаннейшее дело', 'запутаннейшего дела'
'дело           неисправимый' -> мера: 4.901907188094263, кол-во:   1. примеры: 'дела
неисправимого'
'небесчестный   дело    ' -> мера: 4.901907188094263, кол-во:   1. примеры:
'небесчестного дела'
'головоломный   дело    ' -> мера: 4.901907188094263, кол-во:   1. примеры:
'головоломного дела'
```

Можно заметить, что эти словосочетания употребляются буквально по одному разу. Это произошло потому что в тексте есть слова, которые употребляются крайне редко, но всегда вместе со словом дело (например головоломное, запутаннейшее и тд.), что является важным фактором при вычислении Mi-score.

Ищем самые употребляемые словосочетания

Теперь попробуем найти самые употребляемые словосочетания в нашем тексте по метрикам Dice и Mi.

Dice:

```
'укокошить      неантихрист ' -> мера: 1.00000, кол-во:   1, примеры: 'укокошили
неантихристов'
'еше            куриться    ' -> мера: 1.00000, кол-во:   1, примеры: 'еше курилась'
'полмиллиона    сидней      ' -> мера: 1.00000, кол-во:   1, примеры: 'полмиллиона
сидней'
'штаб-офицер    брандер     ' -> мера: 1.00000, кол-во:   1, примеры: 'штаб-офицеров
брандеров'
'двухаршинный   стерлядь    ' -> мера: 1.00000, кол-во:   1, примеры: 'двухаршинными
стерлядями'
'покрытие       бренный     ' -> мера: 1.00000, кол-во:   1, примеры: 'покрытия
бренного'
'федосея        федосей     ' -> мера: 1.00000, кол-во:   1, примеры: 'федосей
федосеевич'
'понукание      нацеплять   ' -> мера: 1.00000, кол-во:   1, примеры: 'понуканьях
нацепляя'
'незапамятный   младенчество' -> мера: 1.00000, кол-во:   1, примеры: 'незапамятного
младенчества'
'утончённейший  гастроном   ' -> мера: 1.00000, кол-во:   1, примеры: 'утонченнейшего
гастронома'
'сомовий        плёс        ' -> мера: 1.00000, кол-во:   3, примеры: 'сомовий плёс'
```

Довольно хорошо видно, что слова из этих биграмм встречаются только вместе (это исходит из того, что мера=1, а значит числитель меры равен знаменателю), поэтому словосочетания попали в лидеры, даже если в тексте встречаются только 1 раз.

Mi:

```
'укокошить      неантихрист ' -> мера: 13.56724, кол-во:   1, примеры: 'укокошили
неантихристов'
'еше            куриться    ' -> мера: 13.56724, кол-во:   1, примеры: 'еше курилась'
'полмиллиона    сидней      ' -> мера: 13.56724, кол-во:   1, примеры: 'полмиллиона
сидней'
'штаб-офицер    брандер     ' -> мера: 13.56724, кол-во:   1, примеры: 'штаб-офицеров
брандеров'
'двухаршинный   стерлядь    ' -> мера: 13.56724, кол-во:   1, примеры: 'двухаршинными
стерлядями'
'покрытие       бренный     ' -> мера: 13.56724, кол-во:   1, примеры: 'покрытия
бренного'
'федосея        федосей     ' -> мера: 13.56724, кол-во:   1, примеры: 'федосей
федосеевич'
'понукание      нацеплять   ' -> мера: 13.56724, кол-во:   1, примеры: 'понуканьях
нацепляя'
'незапамятный   младенчество' -> мера: 13.56724, кол-во:   1, примеры: 'незапамятного
младенчества'
'утончённейший  гастроном   ' -> мера: 13.56724, кол-во:   1, примеры: 'утонченнейшего
гастронома'
```

Результаты аналогичны.

Вывод

Исходя из результатов проделанной работы можно сказать, что с помощью Dice и Mi метрик можно избавиться от биграмм, содержащих союзы, частицы и тд. Также можно заметить, что для Dice метрики более характерно отдавать предпочтение словам, употребляющимся только вместе , в то время как для Mi достаточно уникальности хотя бы одного из слов, чтобы получить высокую оценку. К сожалению оба метода довольно легко зашумляются уникальными словами, что, впрочем, может быть исправлено, если использовать более объемный текст или коллекцию текстов.