### Московский авиационный институт

**Факультет прикладной математики и физики**

**Лабораторная работа №5**

**по курсу:**

**«Обработка естественно-языковых текстов»**

**по теме:**

**«Поиск коллокаций»**

**2 семестр**

Студент: Ахмед С. Х.

Преподаватель: Калинин А. Л.

Группа: 8О-106М

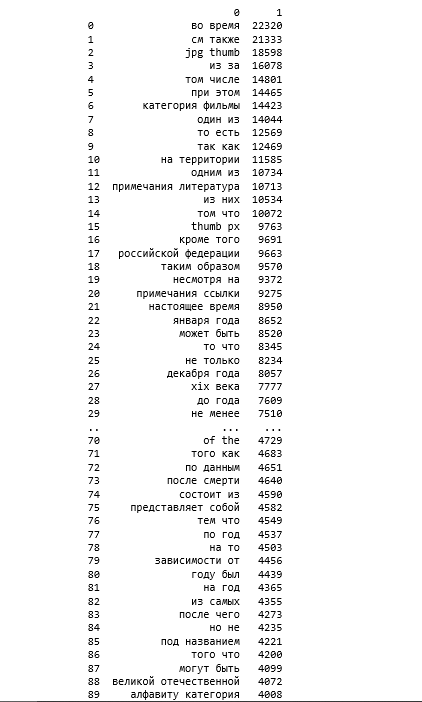
**Москва, 2019 г**

Постановка задачи

Необходимо найти коллокации в имеющемся корпусе, использованного для построения поисковой системы (или его случайному подмножеству достаточного размера). Для поиска коллокаций необходимо использовать как минимум два статистических алгоритма из рассмотренных на лекциях. Сравнить выделенные коллокации между собой, пояснить различия с точки зрения использованых алгоритмов. В отчёте должны быть приведены количество найденных коллокаций, оценка точности метода, примеры найденных коллокаций и ошибочно найденных словосочетаний.

Ход работы

Для поиска коллокаций было решено использовать метод поиска по биграммам(частотность) и расчет взаимной информации. Было решено для данной лабораторной работы воспользоваться половиной своего корпуса для осуществления работы. Поиск по биграммам заключается в том, что будет подсчитано количество всех словосочетаний и наиболее частовстречаемые будут признаны коллокациями. Поиск коллокаций по частным словосочетаниям признается не самым эффективным способом, но все равно его стоит проверить и оценить его эффективность.



Полученный список заполнен случайными фразами, которые нельзя отнести к коллокациям. Однако можно найти коллокации и тут: “великой отечественной”, “российской федерации”,”таким образом”,”несмотря на”.

При расчете взаимной информации обнаружился недостаток метода: если словосочетание встречается 1 раз, то величина взаимной информации будет максимальной. Начало списка найденных коллокаций (у всех словосочетаний одинаковый уровень взаимной информации):





Среди них можно увидеть имена собственные, какие-то случайные номерные названия. Однако можно увидеть и коллокации: “глаз осьминога”,”аркадная платорфма”,”великая сфера”,”вклад спасение” и так далее. Стоит заметить, что высокочастотные термины данный алгоритм обрабатывает уверенно:



Как мы видим это коллокациями не является, что верно и отображено в оценке MI теста.

Выводы

Оба примененных метода имеют свои недостатки: при простом подсчете частот большая вероятность ошибочного определения коллокации из-за того, что существует большое число слов, которые просто часто употребляются друг с другом. Также на частотность тех или иных словосочетаний влияет стиль текстов: новости и художественные произведения имеют разные наборы слов и свои устойчивые словосочетания.

Метод подсчета взаимной информации показал неэффективность в очень редких словосочетаниях: если каждое слово встретилось по 1 разу и только друг с другом (как это часто бывает с редкими именами и фамилиями), то по этому алгоритму, найденное словосочетание будет коллокацией.