**Задание:**

Необходимо выбрать корпус с текстом объемом не менее 10 000 словоупотреблений (например фрагмент корпуса aranea). Посчитать 3-граммы для выбранного корпуса.

1. На основе первого практического задания (далее - ПР1) сформировать данные (список) для расчета n-грамм по словоформам / лексемам (леммами) (в зависимости от доставшегося варианта). Т.е. в варианте со словоформами можно использовать код из пункта 1 в ПР1, а в случае с лексемами можно использовать код из пункта 3 в ПР1.
2. Очистить полученные данные от знаков пунктуации. Можно использовать регулярное выражение: [^\P{P}-]+
3. Привести полученные данные к нижнему регистру.
4. Очистить полученные данные от стоп слов. Можно использовать nltk.corpus stopwords.
5. Посчитать 3-граммы и их частоту.
6. Реализовать алгоритм расчета меры ассоциации (в зависимости от доставшегося варианта)
   1. словоформы, MI
   2. словоформы, t-score
   3. словоформы, log-likelihood \*
   4. лексемы, MI
   5. лексемы, t-score
   6. лексемы, log-likelihood \*

\* повышенный коэффициент

1. Проверить результаты  с помощью [библиотеки](https://www.nltk.org/howto/collocations.html) NLTK, пример скрипта (без удаления стоп слов и пунктуации):

import nltk

from nltk.collocations import \*

from nltk.corpus import PlaintextCorpusReader

bigram\_measures = nltk.collocations.BigramAssocMeasures()

trigram\_measures = nltk.collocations.TrigramAssocMeasures()

f = open('text.txt')

raw = f.read()

tokens = nltk.word\_tokenize(raw,'russian',True)

#print(tokens[:10])

text = nltk.Text(tokens)

#http://www.nltk.org/\_modules/nltk/collocations.html

finder\_bi = BigramCollocationFinder.from\_words(text)

finder\_thr = TrigramCollocationFinder.from\_words(text)

print(finder\_bi.nbest(bigram\_measures.pmi, 10))

print(finder\_thr.nbest(trigram\_measures.pmi, 10))

На защите лабораторной продемонстрировать работу алгоритмов (собственный, nltk), показать список из top-30 3-грамм, объяснить различия в результатах.

**Примечание:**

Подробное описание мер ассоциативной связанности можно изучить в работе [М.В. Хохловой “Экспериментальная проверка методов выделения коллокаций”](https://blogs.helsinki.fi/slavica-helsingiensia/files/2019/11/sh34-21.pdf).

Меры ассоциативной связанности (association measures) - меры, вычисляющие силу связи между элементами в составе коллокации (параметры: частота совместной встречаемости, частота слова в корпусе, размер корпуса, и др.). MI (Mutual Information) чувствительна к низкочастотным словам, а t-score полезна для нахождения высочастотных коллокаций.

MI(n,c)=log2f(n,c) Nf(n) f(c)

где n — ключевое слово (node); c — коллокат (collocate); f (n, c) — частота встречаемости ключевого слова n в паре с коллокатом *c*; f(n), f(c) — абсолютные (независимые) частоты ключевого слова n и слова c в корпусе (тексте); N — общее число словоупотреблений в корпусе (тексте).

log − likelihood = 2ijOij+logOijEij

где Oij , Eij - наблюдаемая и ожидаемая частоты. Рекомендуется более подробно прочитать стр. 346 в работе [М.В. Хохловой “Экспериментальная проверка методов выделения коллокаций”](https://blogs.helsinki.fi/slavica-helsingiensia/files/2019/11/sh34-21.pdf).

t-score = f(n,c)-f(n)f(c)Nf(n,c)

где n — ключевое слово (node); *c* — коллокат (collocate); f(n, c) — частота встречаемости ключевого слова n в паре с коллокатом ; f(n), f(c) — абсолютные (независимые) частоты ключевого слова n и слова *c* в корпусе (тексте); N — общее число словоупотреблений в корпусе (тексте).