Лабораторная работа

Работа с библиотекой Natural Language Toolkit (NLTK) для решения задач

Статистической обработки текстов (английский язык)

Изучение материала на основе документации по NLTK:

<http://www.nltk.org/>

Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit

Steven Bird, Ewan Klein, and Edward Loper,

<http://www.nltk.org/book_1ed/>

Эксперименты проводить в <https://colab.research.google.com/>

Ход выполнения работы

Задание для выполнения:

1 – этап экспериментов (Анализ структуры отдельного корпуса (текста))

1. Выбрать корпус для экспериментов (из NLTK или загрузить свой)
2. Провести статистический анализ текста:
* Длина текста, словарь текста, число различных слов в словаре, рассчитать параметр лексического разнообразия текста.
* Определить число предложений, слов (провести токенизацию).
* Убрать стоп слова (предлоги, союзы, управляющие слова) и построить частотный график встречаемости слов в тексте. Кумулятивный график частотного распределения слов.
* Выделить частотные слова, относящиеся к одной леме (провести лематизацию)
* На основе результатов лематизации вывести на печать слова, определяющие тематику текста (претенденты на ключевые слова). Выделить по частоте и длине.
* Провести исследование тематической структуры текста (в каких частях текста о чем говориться) – исследовать частотное *расположение* слов в тексте - построить*график дисперсии.*
* Распечатать ключевые слова (частотные слова), относящиеся к наиболее тематически важному разделу текста (определить по графику дисперсии). Для них построить частотный график встречаемости слов в тексте. Кумулятивный график частотного распределения слов.
* Для ключевых слов найти им соответствующие биграммы и коллокации в тексте, оценить их частотность. Экспертным методом проверить соответствуют ли определенные словосочетания важными для уточнения тематики текста.

2 – этап экспериментов (сравнительный анализ нескольких корпусов)

1. Выбрать 2-3 корпуса для экспериментов (из NLTK или загрузить свои)
2. Провести статистический анализ этих корпусов по плану задания 1.
* Провести сравнение по статистическим параметрам: словарь текста, число различных слов в словаре, рассчитать параметр лексического разнообразия те (насколько ) текста.
* Исследовать тематические структуры текстов (в каких частях текстов о чем говориться) – исследовать частотное *расположение* слов в тексте - построить*график дисперсии.*

Отчетность по лабораторной работе

* По проведенным исследованиям сделать отчет с заключением о статистике, стилистике, тематике исследуемых текстов.
* Отчет оформляется в Word с описанием проведенного исследования

с питоновскими скриптами и комментариями в блокноте, разработанном в <https://colab.research.google.com/>.

* Дайте название своему проекту с экспериментами (щелкнув на имя блокнота в правом верхнем углу, переименуйте). Все ваши эксперименты сохраняются на вашем гугл диске.
* В отчет нужно прикрепить ссылку на созданный блокнот с экспериментами – *Поделиться* (в правом верхнем углу colab.research). Отчет выложить в ответах на задание на мудле.