**Домашнее задание №2.**

**Корпусная лингвистика, статистика, языковые модели**

На выбор варианты A, B, C, D, Е (*в скобках указаны баллы оценки)*

**Срок выполнения задания** – до **12 марта** включительно.

**Отчет высылается** **по почте** Олегу Дону на адрес [dragn613@yandex.ru](file:///C:\Users\Elena\Documents\Елена\ВШЭ\Курс_КЛ\Курс_КЛ_2023\ДЗ_КЛ_2023\dragn613@yandex.ru)

В качестве темы письма указать: «**КЛ ДЗ № 2, ФИО**»

***Замечания:***

1. Оценка задания учитывает объем и качество проведенной работы, включая подробность отчета. Более полные решения, подробный анализ результатов, итоговые выводы обычно дают дополнительные баллы.
2. Отчет д. быть в виде обычного текста, код программы присылается отдельным файлом.

**A.** *(5-9 баллов)*

Выбрав 1-2 слова из следующего списка, исследовать временные изменения их смысла и частоты употребления, а также другие особенности употребления:

*беспредел, бомбить, ватрушка, грузить, достать, капуста, кидать, клеить, крутой, корзина, лимон, мочить, наварить, наехать, наколоть, обуть, откат, продвинутый, пятачок, совок, тащиться, тупить, упакованный, фишка, шестерка, чайник.*

Для этого:

1) Найти и сравнить значения и толкования слова в словарях Даля, Ушакова, Ожегова, Кузнецова (см. <http://dic.academic.ru/>)

2) Исследовать употребление этого слова в НКРЯ ([www.ruscorpora.ru](http://www.ruscorpora.ru)):

– в основном корпусе

– в газетном корпусе (рассмотреть подкорпуса для разных периодов времени)

3) Проанализировать график употребления слова (https://books.google.com/ngrams)

4) Посмотреть использование слова на **ya.ru**

5) При возможности (и по желанию) – исследовать употребление слова в корпусе RuTenTen11 системы *SketchEngine* (https://www.sketchengine.co.uk), предварительно зарегистрировавшись в ней (если получится).

Отчет: Описание сходства и различий смысла выбранного слова/слов в разных словарях, а также в реальном употреблении (корпуса, яндекс-тексты) – все ли значения слова есть в словарях? ; описание изменений в частоте употребления, краткая характеристика использованных словарных и поисковых ресурсов; выводы.

**В.** *(6-10 баллов)*

Найти в Интернете уже вычисленную статистику N-грамм для любого европейского языка (например, взяв для русского языка данные проекта *ОpenСorpora*  <http://opencorpora.org/?page=downloads>)

и рассчитать (вручную или составив для этого программу) вероятности:

– 5-7 фраз, покрываемых этой моделью;

– 5-7 фраз, не покрываемых ею, используя при этом один из методов сглаживания (или сравнив два метода сглаживания).

Рассчитать также перплексию – коэффициент неопределенности построенной модели, выбрав для этого соответствующий текст.

Если взяты две N-граммные модели: для N=2 и N=3, рассчитать вероятности и сравнить точность предсказания слов и перплексию для этих моделей.

Отчет: Краткое описание используемой статистики (базы N-грамм), исследуемые фразы и текст для оценки перплексии, проделанные расчеты и вычисленные оценки, программа (если составлялась), сравнения и выводы.

**С.**  *(7-11 баллов)*

Программно построить N-граммную модель языка (N=1,2) по выбранной коллекции текстов (например, из Интернета), написав свою программу или воспользовавшись доступными средствами/инструментами.

Рассчитать по этим моделям и сравнить вероятности:

– 5 фраз, присутствующих в исходной текстовой коллекции

– 5 фраз, отсутствующих в ней, используя при этом один из методов сглаживания (или сравнив два метода сглаживания).

Рассчитать также перплексию – коэффициент неопределенности построенной модели выбрав для этого соответствующий текст.

Если построены две N-граммные модели, рассчитать вероятности и сравнить точность предсказания слов и перплексию для этих моделей.

Отчет: Характеристика исходной текстовой коллекции (в том числе, как/откуда она получена), описание построенной языковой модели/ей и метода построения, проделанные расчеты и вычисленные оценки и сравнения, программа с комментариями, выводы.

**D.**  *(7-11 баллов)*

Провести исследование явления "разреженности данных" в коллекциях/корпусах текстов, выбрав для этого самостоятельно текст большого объема или коллекцию текстов (например, объединив несколько текстов из библиотеки Машкова). Определить общее количество и процент отсутствующих в выбранном тексте словарных словоформ и лемм, используя для этого один из доступных словарей языка, например, словарь OpenCorpora (<http://www.opencorpora.org>) или частотный словарь <http://dict.ruslang.ru/freq.php>.

Определить также наиболее частотную часть речи отсутствующих словоформ/лемм, число и примеры отсутствующих лемм различных частей речи и другую статистику разреженности.

Отчет: краткая характеристика исследуемого текста или текстовой коллекции, описание способа оценки разреженности данных и вычисленные оценки разреженности, указание примененных средств лемматизации, примеры отсутствующих в тексте/коллекции лемм, текст составленной программы, выводы по проведенному исследованию.

**Е.**  *(7-11 баллов)*

Провести анализ того, как явление "разреженности данных" в коллекциях/корпусах текстов отражается в предобученных предсказательных языковых моделях *Word2Vec* и/или *GloVe* (но не *FastText*). Модели могут быть взяты, например, из библиотеки *Gensim* (<https://radimrehurek.com/gensim/models/word2vec.html>) или из ресурса *RusVectores* (<http://rusvectores.org/ru/models/> ). Для выбранных языковых моделей подсчитать общее количество и процент отсутствующих в них векторов слов (ембеддингов), используя для этого один из доступных словарей языка: словарь словоформ/лемм OpenCorpora <http://www.opencorpora.org> или частотный словарь <http://dict.ruslang.ru/freq.php> или др.

Определить также наиболее частотную часть речи отсутствующих словоформ/лемм, число и примеры отсутствующих лемм различных частей речи и другую статистику разреженности.

Отчет: краткая характеристика исследуемых языковых моделей (включая текстовую коллекцию, на которой она обучалась), указание примененных при исследовании средств и ресурсов, описание способа оценки разреженности данных и вычисленные оценки разреженности, примеры отсутствующих векторов слов, текст составленной программы с комментариями, выводы по проведенному исследованию.