

Домашнее задание №2

Дудырев Егор

10 марта 2020

1 Задание

Задание Е

Программно построить по выбранной коллекции текстов две N-граммные модели - для $N=2$ и $N=3$, используя доступные средства (например, см. вариант D). Рассчитать по этим моделям и сравнить вероятности:

1. 5 фраз, присутствующих в исходной текстовой коллекции
2. 5 фраз, отсутствующих в ней, используя при этом один из методов сглаживания

Рассчитать также перплексию - коэффициент неопределённости построенных моделей. Оценить и сравнить точность предсказания слов из указанных отсутствующих фраз построенными моделями.

Отчёт: Характеристика исходной текстовой коллекции (в том числе, как/откуда она получена), описание построенной языковой модели и метода её построения, проделанные расчёты, программа с комментариями, выводы.

2 Текстовая коллекция

В качестве коллекции текстов взяты 5 первых тома Полного собрания сочинений Л.Н. Толстого, а именно:

- Том 1. Детство
- Том 2. Юность
- Том 3. Произведения 1852-1856 гг.
- Том 4. Произведения севастопольского периода. Утро помещика

- Том 5. Произведения 1856-1859

Томы 1-4 использовались для расчёта N-грамм. Том 5 использовался для расчёта перплексий. В текстах, помимо собственно художественного произведения, содержится также различная побочная информация - ссылки, комментарии редакции и т.п.

Количество слов, на которых строились N-граммы - 346790.

Тестовые фразы:

- Фразы из текста

1. “Матушка сидела в гостиной и разливала чай”, Том1 детство
2. “И он ударил вилкой по столу”, Том 1 детство
3. “в карты не играл, кутил редко и курил простой табак”, Том 3. Произведения 1852
4. “Длинные чистые сакли с плоскими земляными крышами и красивыми трубами были расположены по неровным каменистым буграм, между которыми текла небольшая река”, Том 3. Произведения 1852
5. “Не видя никого в избе, Нехлюдов хотел уже выйти, как протяжный, влажный вздох изобличил хозяина” Том4 Произведения севастопольского периода. утро помещика

- Фразы не из текста

1. “Но вот где является в полном блеске историческое воззрение г. Маркова”, Том 8. Педагогические статьи 1860
2. “Что же это такое понятие прогресса и вера в него”, Том8. Педагогические статьи 1860
3. “Высунувшееся из кареты лицо Наташи сияло насмешливою ласкою”, Том9. Война и мир
4. “Наполеон испытывал то несколько завистливое и беспокойное любопытство, которое испытывают люди при виде форм не знающей о них, чуждой жизни”, Том9. Война и мир
5. “Несколько купцов столпились около офицера”, Том9. Война и мир

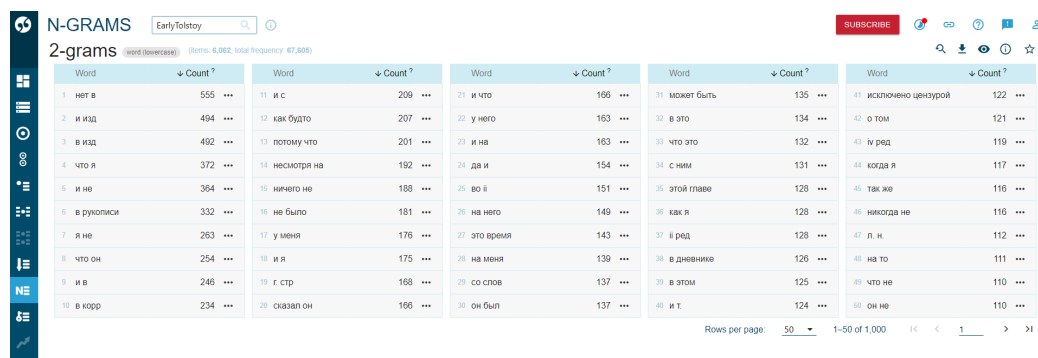
3 Модели Sketch Engine

Изначально для построения N-грамм был использован сервис Sketch Engine. Настройки и для 2N-грамм и для 3N-грамм следующие:

1. Отсутствие разделения на печатные и строчные символы
2. Атрибут - слово
3. Минимальная частота - 0
4. Остальные значения по умолчанию

Полученные модели указаны на рис. 1 и рис. 2.

Рис. 1: Биграммы, полученные с помощью Sketch Engine



Word	Count	Word	Count	Word	Count	Word	Count	Word	Count
нет в	555	и с	209	и что	166	может быть	135	исключено цензурой	122
и изд	494	как будто	207	у него	163	в это	134	о том	121
в изд	492	потому что	201	и на	163	что это	132	и ред	119
что я	372	несмотря на	192	да и	154	с ним	131	куда я	117
и не	364	ничего не	188	во и	151	этой главе	128	так же	116
в рукописи	332	не было	181	на него	149	как я	128	никогда не	116
я не	263	у меня	176	это время	143	и ред	128	л. н.	112
и он	254	и я	175	на меня	139	в дневнике	126	на то	111
и в	246	к стр	168	со слов	137	в этом	125	что не	110
в корр	234	сказал он	166	он был	137	и т.	124	он не	110

Рис. 2: Триграммы, полученные с помощью Sketch Engine



Word	Count	Word	Count	Word	Count	Word	Count
в это время	115	редакции этой главе	48	не красная строка	34	главы iv ред	24
этой главе соответствует	95	г. исключено цензурой	48	в и ред	34	в эту минуту	24
в рукописи и	92	г. редакции	44	я не могу	33	со всех сторон	23
и т. д.	87	близок к тексту	43	этой главы нет	33	не может быть	23
нет в изд	82	не только не	41	редакции этой главы	33	л. н. толстого	23
несмотря на то	77	редакции этой	41	по крайней мере	32	ежели бы я	23
во и ред	76	редакции этой	40	в архиве толстого	32	в дневнике под	21
в то время	76	comme il faut	39	тексту окончательной редакции	30	с тех пор	20
л. н. толстого	74	в первый раз	38	вместе с тем	30	но потом исправлено	20
во и редакции	69	в и ред	36	я не мог	29	ни за что	20
рукописи и изд	58	на другой день	35	что я не	29	и вместе с	20
нет в рукописи	55	и т. п.	35	к тексту окончательной	29		
нет в корр	52	роман русского помещика	34	на следующий день	24		

4 Собственные модели

После расчёта моделей Sketch Engine оказалось, что даже при заданном параметре "Минимальная частота - 0 настоящая минимальная частота N-грамм в получившихся данных равна 3. Т.е., по каким-то причинам, Sketch Engine не сохранил редко употребляемые N-граммы, поэтому модель без сглаживания может указать вероятность фразы 0 даже на предложениях из обучающей текстовой коллекции.

В результате, N-граммы были рассчитаны собственноручно на основе тех же данных, которые подавались в Sketch Engine.

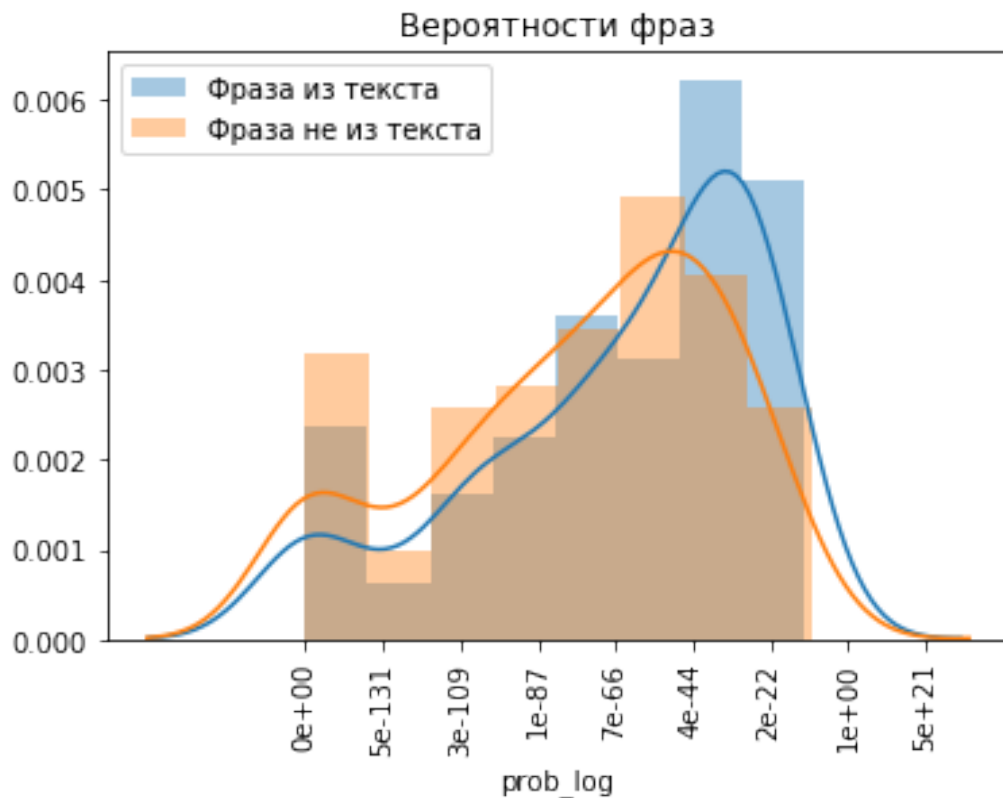
5 Результаты

Для получения результатов рассчитаем вероятности фраз всеми имеющимися моделями (2-3 N-граммы, Sketch Engine и собственные модели) для 10 равномерно распределённых (в логарифмическом масштабе) значений параметра α .

Получим следующие зависимости:

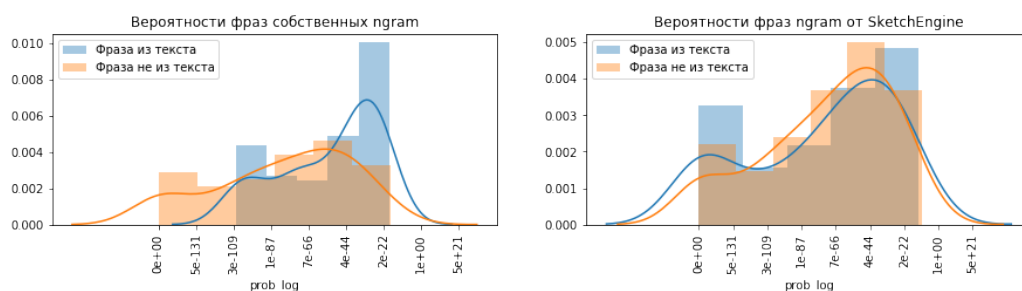
1. Рассчитанные вероятности для фраз из обучающего текста в целом выше, чем для фраз которых в тексте не было (рис. 3).

Рис. 3: Вероятности обучающих и тестовых фраз



2. Модели, учитывающие самые редкие N-граммы показывают сильнее разделяют фразы которые были в обучающем наборе от тех, которых не было (рис. 4).

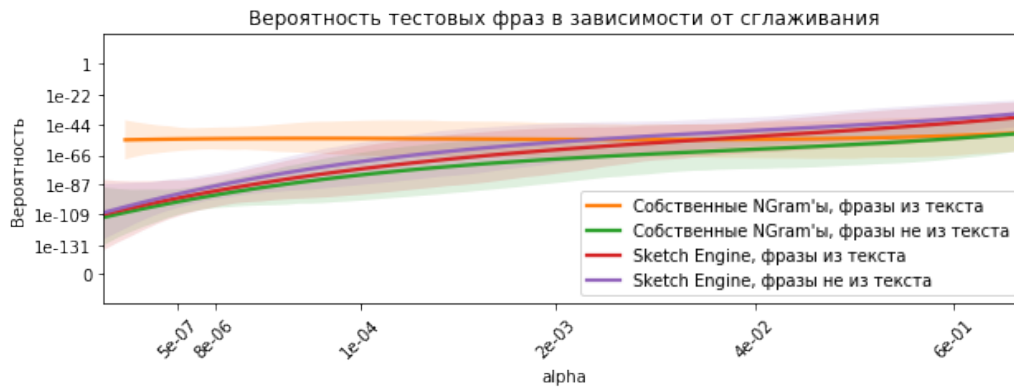
Рис. 4: Вероятности обучающих и тестовых фраз в разрезе моделей



3. При повышении коэффициента сглаживания вероятности фраз, которых не было в обучающем корпусе, повышаются. Фразы, которые

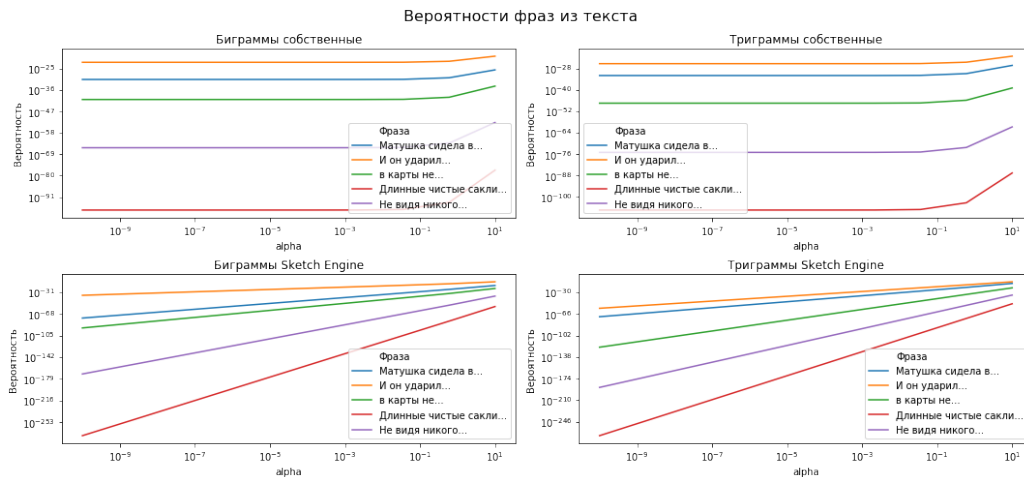
были в корпусе, показывают стабильно высокую вероятность (рис. 5).

Рис. 5: Вероятности фраз при изменении коэффициента сглаживания



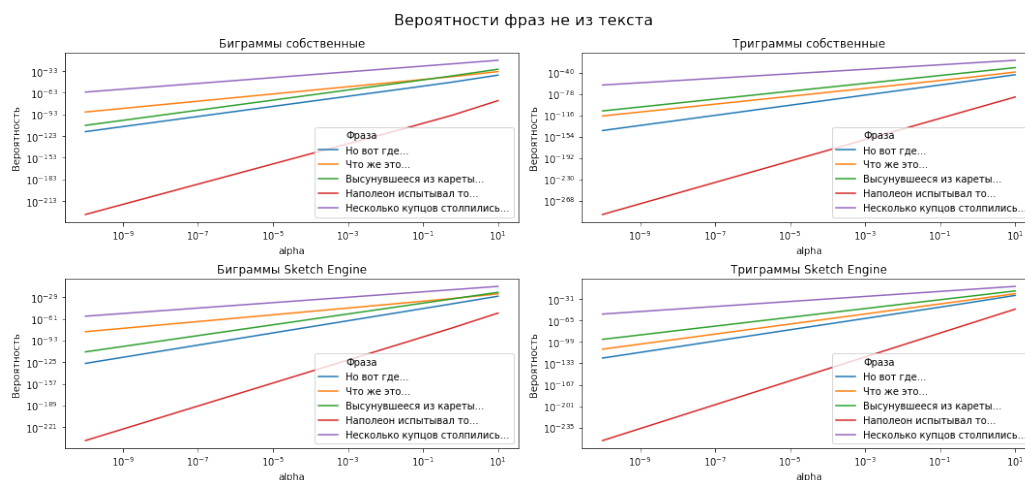
4. Рассмотрим только фразы из обучающего корпуса. Для моделей со всеми N-граммами вероятности фраз стабильно высокие. Для моделей без наиболее редких N-грамм вероятности возрастают с увеличением параметра сглаживания (рис. 6).

Рис. 6: Вероятности фраз из обучающей коллекции текстов



5. Для фраз не из обучающего корпуса вероятности повышаются с увеличением параметра сглаживания для всех моделей (Рис. 7).

Рис. 7: Вероятности фраз не из обучающей коллекции текстов



Для расчёта перплексии моделей использовались предложения из пятого тома Полного собрания сочинения Л.Н. Толстого. Слова с 1000 по 1050. Предложения большей длины часто имели вероятность 0, что давало бесконечную перплексию.

Рис. 8: Перплексии моделей



Вывод: лучше всего для расчёта вероятности текста использовать все возможные N-граммы текста, даже если они встречаются всего один раз. С другой стороны, это может быть накладно с точки зрения используемых ресурсов. При невозможности хранить все возможные N-граммы необходимо использовать сглаживание со значением alpha близким к 1.

6 Ссылки на код

- Исполняемый код