МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«ДОНСКОЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
(ДГТУ)

Отчет по лабораторной работе №1

Исследование применимости законов Зипфа к русскоязычным текстам

Выполнил:
студент МИН21
Урывский Д.В.

Ростов-на-Дону
2020

Цель работы

В ходе лабораторной работы получить практические навыки морфологического анализа текста, применимости законов Зипфа к русскоязычным документам и оптимизации поиска соответствующей информации в Интернет.

Ход выполнения

Определим частоту вхождения слов (ссылка на статью http://itno.e.donstu.ru/documents/articles/313-316.pdf).

Таблица 1. Частота вхождения слов

Ранг	Частота	Слово
1	10	система
2	7	студент
2	7	изучение
3	6	курс
4	4	работа
5	3	направление
5	3	механики
5	3	механика
5	3	язык
5	3	прикладной
5	3	подготовки
5	3	отдельный
5	3	процесс
5	3	позволять
5	3	компьютерных
5	3	компьютерной
5	3	комплекс

Определим вероятность вхождения произвольно выбранного слова в текст. Очевидно, она будет равна отношению частоты вхождения этого слова к общему числу слов в тексте. Таким образом, справедливо следующее выражение:

Вероятность = Частота вхождения слова / Число слов

Вероятность слова “изучение” = 7 / 700 = 0,01 = 1%

Если умножить вероятность обнаружения слова в тексте на ранг частоты, то получившаяся величина (С) – константа Зипфа приблизительно постоянна:

С = (Частота вхождения слова х Ранг частоты) / Число слов

С = (“изучение” 7 * 2) / 700 = 0,02

С = (“курс” 6 * 3) / 700 = 0,026

С = (“язык” 5 * 3) / 700 = 0,021

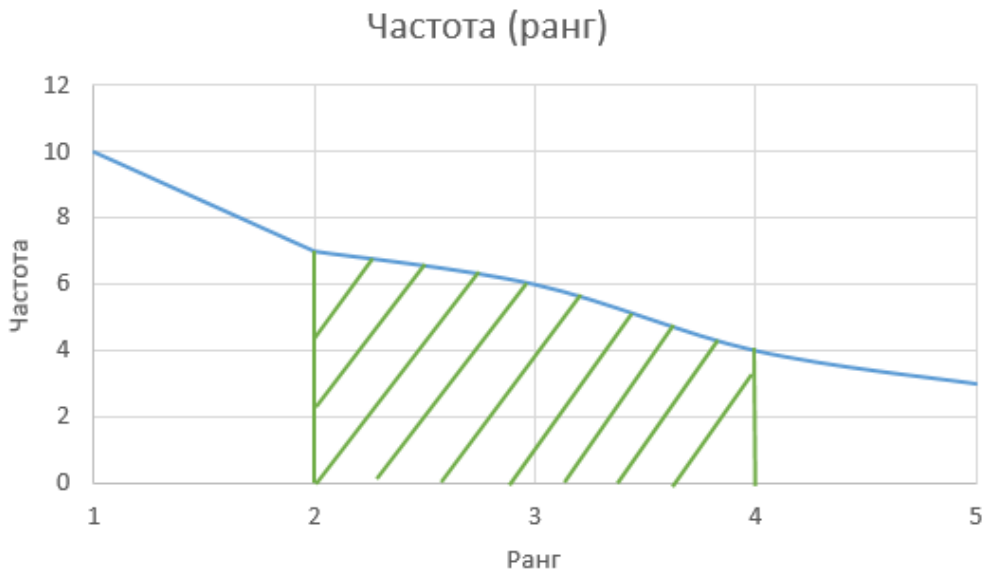

Рис. 1. Диаграмма частота – ранг

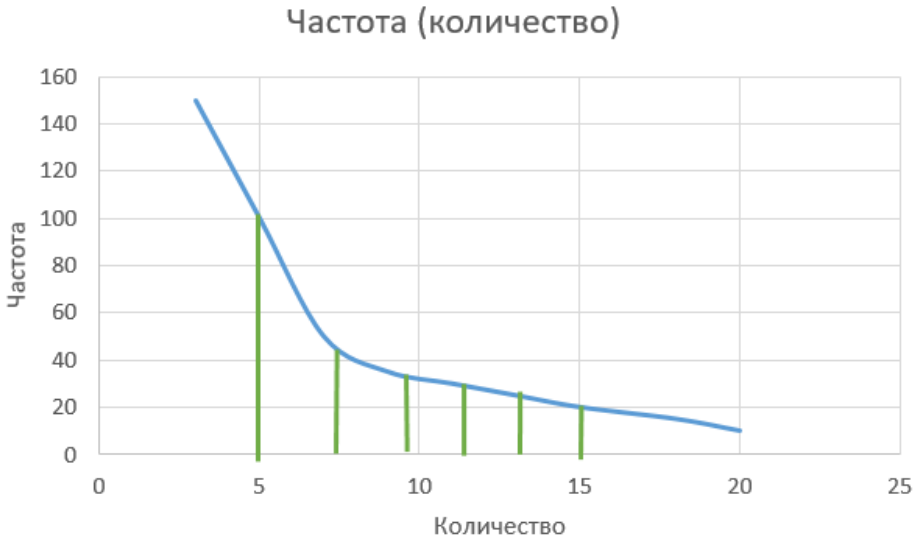

Рис. 2. Диаграмма частота – количество

Исходя из анализа был создан поисковой запрос, состоящий из слов: система студент изучение курс работа направление механики механика язык прикладной.

Контрольные вопросы:

1. Первый закон Зипфа – «ранг - частота». Вероятность обнаружения любого слова, умноженная на его ранг — постоянная величина. В любом тексте, написанном человеком, этот закон статистически верен.
2. Второй закон Зипфа - "количество - частота ". Первый закон не учитывает факт того, что разные слова могут входить в текст с одинаковой частотой. Ципф установил, что частота и количество слов, входящих в текст с этой частотой, также имеют зависимость.
3. Законы Зипфа универсальны. В принципе, они применимы не только к текстам. Аналогичный вид имеет, например, зависимость количества городов от числа проживающих в них жителей. Характеристики популярности сайтов в сети Интернет - тоже отвечают законам Зипфа. Не исключено, что в них отражается "человеческое" происхождение объекта. Рассмотрим другой пример. Хорошо известно, что ученые давно бьются над расшифровкой манускриптов Войнича.
4. Выберем любое слово и посчитаем, сколько раз оно встречается в тексте. Эту величи-ну определим, как частоту вхождения слова и измерим её. Некоторые слова будут иметь оди-наковую частоту, то есть входить в текст равное количество раз. Сгруппируем их, взяв только одно значение из каждой группы. Расположим частоты по мере их убывания и пронумеруем. Порядковый номер частоты называется её рангом. Так, наиболее часто встречающиеся слова будут иметь ранг 1, следующие за ними - 2 и т.д.

Определим вероятность вхождения произвольно выбранного слова в текст. Очевидно, она будет равна отношению частоты вхождения этого слова к общему числу слов в тексте. Таким образом, справедливо следующее выражение:

Вероятность = Частота вхождения слова / Число слов (1)

Зипф обнаружил закономерность - если умножить вероятность обнаружения слова в тексте на ранг частоты, то получившаяся величина (C) – константа Зипфа приблизительно постоянна:

C = (Частота вхождения слова x Ранг частоты) / Число слов (2) текста

5. При рассмотрении первого закона, не учитывался факт, что разные слова могут вхо-дить в текст с одинаковой частотой. Зипф установил, что частота и количество слов, входя-щих в текст с этой частотой, тоже связаны между собой.

Если построить график, отложив по одной оси (оси X) частоту вхождения слова, а по другой (оси Y) - количество слов в данной частоте, то получившаяся кривая будет сохранять свои параметры для всех без

исключения созданных человеком текстов! Как и в предыдущем случае, это утверждение верно в пределах одного языка. Однако и межъязыковые различия невелики. На каком бы языке текст ни был написан, форма кривой Зипфа останется неизмен-ной. Могут немного отличаться лишь коэффициенты, отвечающие за наклон кривой (Рисунок. 1). Следует заметить, что в логарифмическом масштабе, за исключением нескольких начальных точек, график зависимости количества слов от частоты представляет собой пря-мую линию.

6. До сих пор рассматривался отдельно взятый документ, не принимался во внимание тот факт, что он входит в базу данных наряду с множеством других документов. Если представить всю базу данных как единый документ, к ней можно будет применить те же законы, что и к единичному документу. Чтобы избавиться от лишних слов и в тоже время поднять рейтинг значимых слов, вводят инверсную частоту термина. Значение этого параметра тем меньше, чем чаще слово встречается в документах базы данных.

7. Предположим, база данных имеет 8 документов (Д1, Д2, … Д8), в которых содержатся 12 терминов (см. таблицу). Если термин входит в документ, в соответствующей клетке таб-лицы проставляется единица, в противном случае - ноль (в реальной базе поисковой машины все сложнее: помимо прочего, учитываются еще и весовые коэффициенты терминов).

Составим, например, такой запрос: «трубопроводы к сепараторам». Поисковая система обработает запрос: удалит стоп - слова и, возможно, проведет морфологический анализ. Останется два термина: трубопровод и сепаратор. Система будет искать все документы, где встречается хотя бы один из терминов. Посмотрим на матрицу. Пусть указанные в запросе термины есть в документах: Д1, Д2, Д4, Д7, Д8. Они и будут выданы в ответ на запрос. Однако нетрудно заметить, что документы Д4 и Д7 не удовлетворяют нашим запросам - они из области выпечки хлеба и никакого отношения к химико-технологическому оборудованию не имеют. Впрочем, поисковая машина все сделала правильно, ведь, с ее точки зрения, термины трубопровод и сепаратор равноценны

8. Пространственно-векторная модель позволяет получить результат, хорошо согласующийся с запросом. Причем документ может оказаться полезным, даже не имея 100% соответствия. В найденном документе может вовсе не оказаться одного или нескольких слов запроса, но при этом его смысл будет запросу соответствовать.

9. Релева́нтность (англ. relevant) — применительно к результатам работы поисковой машины — степень соответствия запроса и найденного, уместность результата. Это субъективное понятие, поскольку результаты поиска, уместные для одного пользователя, могут быть совершенно неприемлемыми для другого.

10. Пертине́нтность (англ. pertinent) - соотношение объема полезной информации к общему объему полученной информации.

11. Релевантность — смысловое соответствие между информационным запросом и полученным сообщением. Аквариум — разведение рыбок. Пертинентность — точное соответствие полученной информации информационной потребности пользователя. Аквариум — Виктора Суворова. Проще говоря, пертинентность — лучше чем релевантность, потому что точнее.