МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В.Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра алгоритмических языков

Автоматическое определение авторства литературных текстов

Отчет по спецкурсу "Математические методы анализа текста"

Студент: Александр Желубенков
Группа: 325
Преподаватель: Dr Мстислав Масленников

Москва, 2013

Введение

Для определения авторства текстов можно использовать свойства с различной частотностью. Например, служебные слова являются самыми частотными, а n-граммы низкочастотными при больших n. Чтобы эффективно определять авторство, важно понимать, стоит ли использовать высокочастотные или низкочастотные свойства.

В данном отчёте рассматривается проблема оптимальной комбинации высокочастотных (служебных) слов и одних из самых низкочастотных слов (триграмм) для классификации текстов. Также исследуется, какое количество триграмм является оптимальным для определения автора.

Постановка задачи

Дано множество текстов (текстовых файлов). У каждого из текстов определен автор.
Требуется подобрать свойства и натренировать классификатор, распознающий каждого автора.

Свойства текстов

Используем следующие варианты выбора свойств:

1. Частота служебных слов и знаков
2. Частота наиболее часто встречающихся триграмм

1. Служебные слова и знаки

Знаки: ‘:’, ‘;’, ‘…’
Предлоги: -в, -без, -до, -из, -к, -на, -по, -о, -от, -перед, -при, -через, -с, -у, -за, -над, -об, -под, -про, -для;
Союзы, частицы: -и, -или, -как, -словно, -а, -что, -но, -однако, -чтобы, -когда, -бы, -б, -лишь -едва, -точно, -будто, -если, -да, -притом, -же, -ж, -тоже, -либо, -самый, -весь, -очень, -крайне, -ли, -ль, -совсем, -хотя, -чтоб;
Свойства – частота появления выбранных слов, определяющаяся как отношение числа раз, которое слово встретилось в тексте, к общему количеству слов в тексте.

2. Наиболее часто встречающиеся триграммы

В русском языке у одного слова разные окончания в различных падежах, поэтому нам необходимо отождествлять слова с одинаковой основой. Сначала мы обрезаем каждое слово и оставляем только 5 первых букв. Три подряд идущих обрезанных слова мы считаем *триграммой*. При таком подходе слова совпадают, если у них совпадают первые 5 букв.
По каждому автору выбираем K наиболее часто встречающихся триграмм. Здесь частота появления триграммы – отношение количества текстов данного автора, в которых он встречался, к количеству текстов данного автора.
Для N авторов получим K * N триграмм с повторениями. Удалим повторяющиеся триграммы. Оставшиеся M триграмм будут соответствовать M свойствам, где свойство - частота появления данной триграммы в тексте.

Построение классификатора

При построении классификатора используется линейный метод опорных векторов. (http://ru.wikipedia.org/wiki/Метод_опорных_векторов)
Задачу мульти-классовой классификации сводим к задачам бинарной классификации, используя стратегию “one vs others”.

Реализация
На основании данного выбора свойств реализована программа, позволяющая строить классификатор и классифицировать тексты. Используется библиотека LIBLINEAR.

Исходные данные
При тестировании использовались литературные тексты разных авторов, взятые со страницы спецкурса ВМК МГУ «Введение в обработку текстов»
Количество авторов - 20
Тренировочная выборка - trainingSet.zip (128 текстов)
Тестовая выборка - testingSet.zip (84 текста)
Средний размер текста – 200 КБ

Результаты экспериментов
Целью экспериментов является выявление взаимодействия служебных слов и триграмм и нахождение оптимального количества триграмм при выборе свойств автора.
Точность результата определяется, как количество правильно классифицированных текстов из тестовой выборки, к количеству текстов в тестовой выборке.

#Триграмм		Со служебными словами
K = 0	0	0.75
K = 50	0.76	0.76
K = 100	0.83	0.82
K = 200	0.82	0.82
K = 300	0.82	0.82
K = 400	0.82	0.82
K = 500	**0.86**	**0.86**
K = 600	0.83	0.83
K = 700	0.83	0.84
K = 800	0.83	0.83
K = 900	0.82	0.82
K = 1000	0.81	0.81

Выводы:

1. Служебные слова без использования триграмм дают неплохой результат в 75%, но их влияния на точность при использовании триграмм практически нет. Это можно объяснить тем, что уже при K = 50 количество свойств, занимаемых триграммами, равно 575, в то время как служебные слова занимают 55 свойств.
2. Для конкретной тренировочной выборки оптимальное количество триграмм равно 500. При этом результат ухудшается как при слишком большом, так и при слишком малом количестве триграмм.