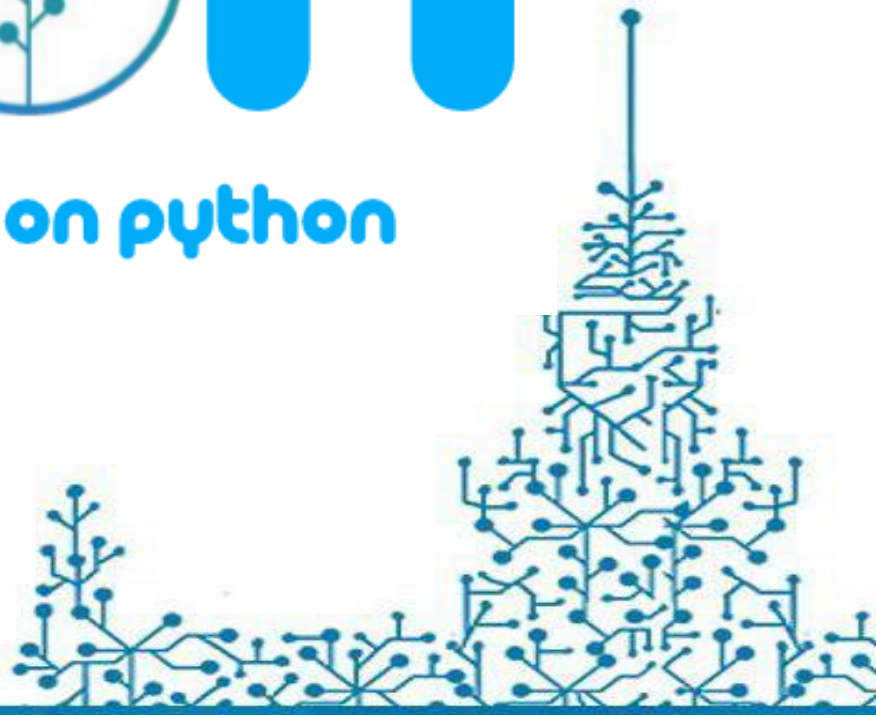
twiton

on python



ЗАДАЧА

- Построить и обучить модель, которая основываясь на тексте твита определяет его тональность (положительная / негативная)

ДАТАСЕТ

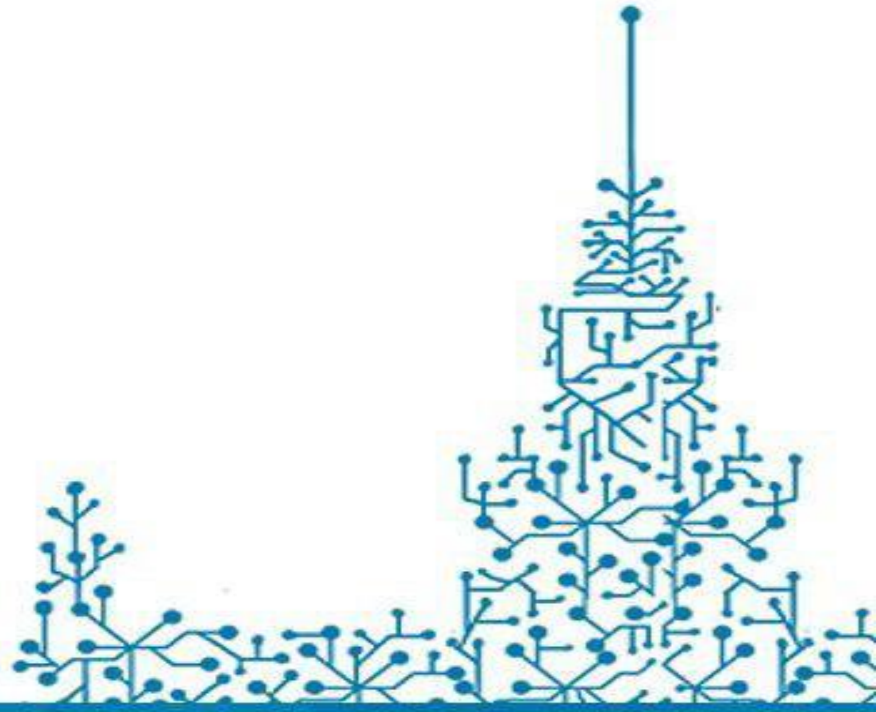
- 100к положительных твитов
- 100к негативных твитов
- дополнительный парсинг сайта twitter.com

ОНЛАЙН АНАЛИЗАТОР

- Сервер с моделью и парсером
- Сайт для взаимодействия пользователя

МЕТРИКА

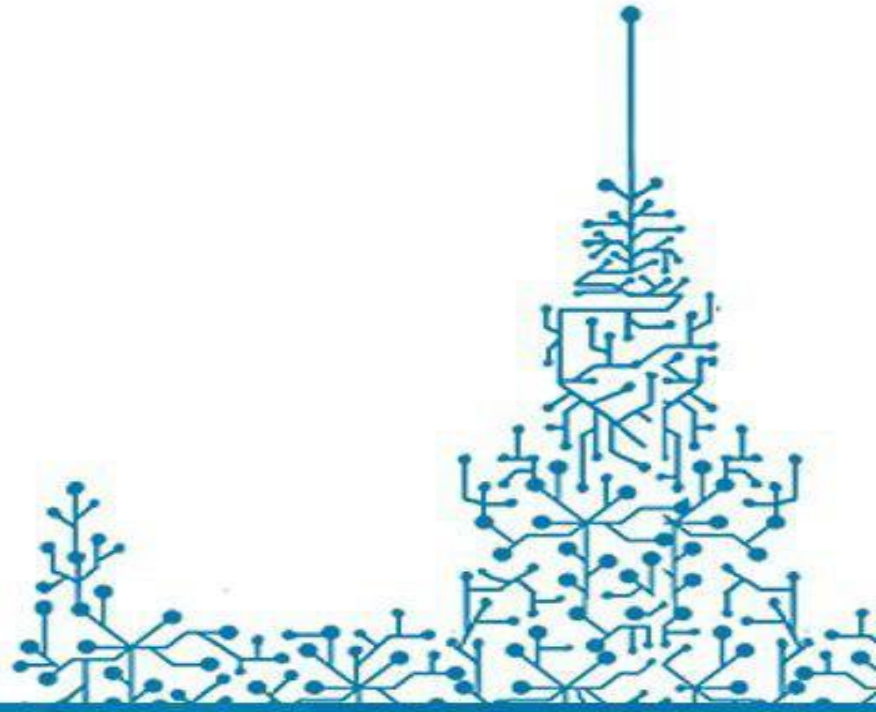
- F1-мера



Обработка данных

- Из текста твитов удаляются:
  - латинские буквы
  - цифры
  - смайлики
  - знаки препинания

- Морфологическая обработка:
  - стемминг (пакет nltk)
  - лемматизация (mystem)

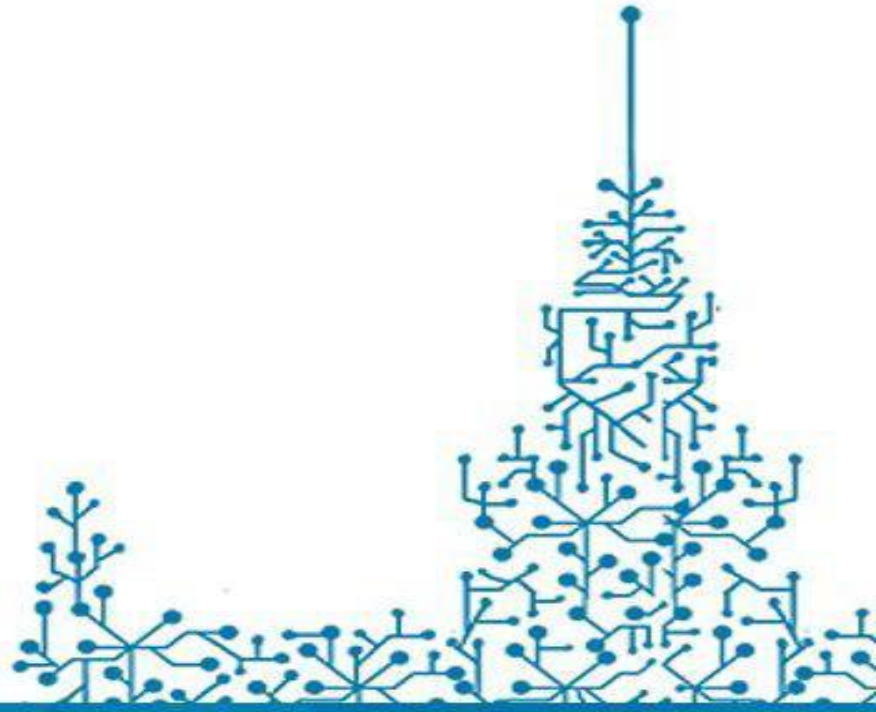


Извлечение признаков

- Count Vectorizer
- TF-IDF Vectorizer
- Word2Vec

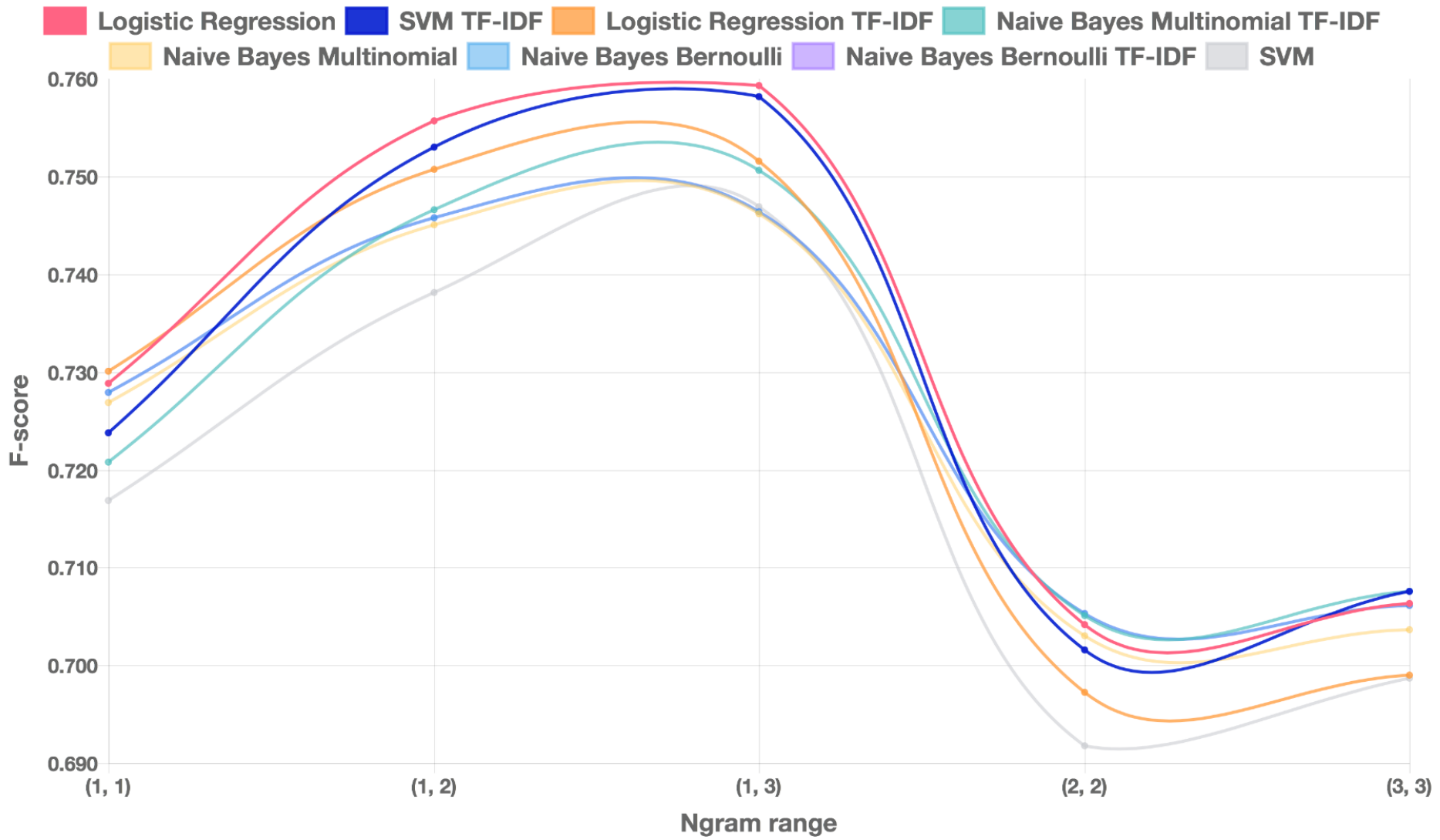
Модели

- Logistic regression
- SVM
- Naive Bayes



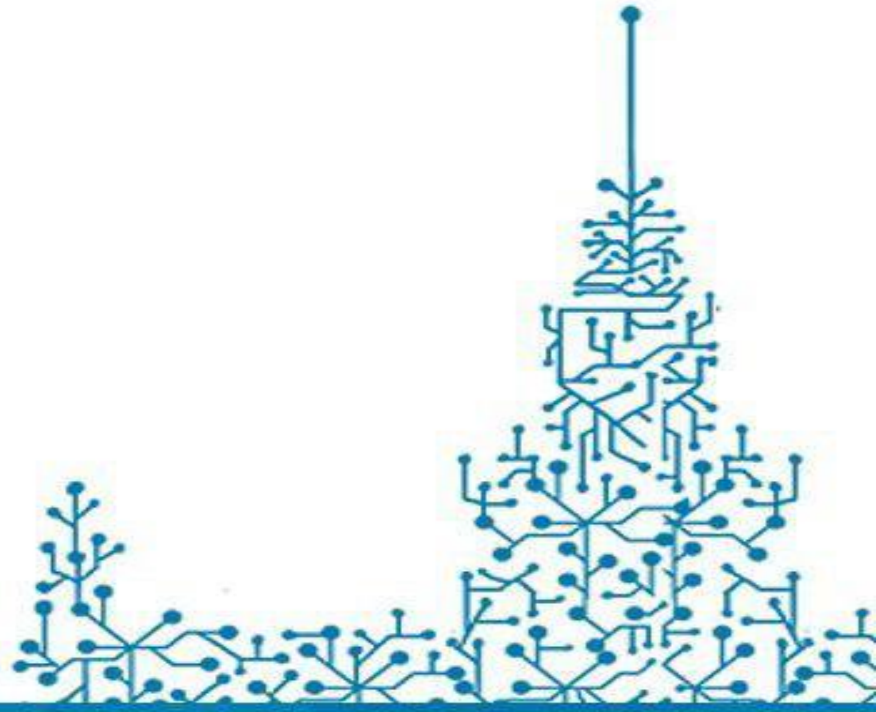
Подбор параметров

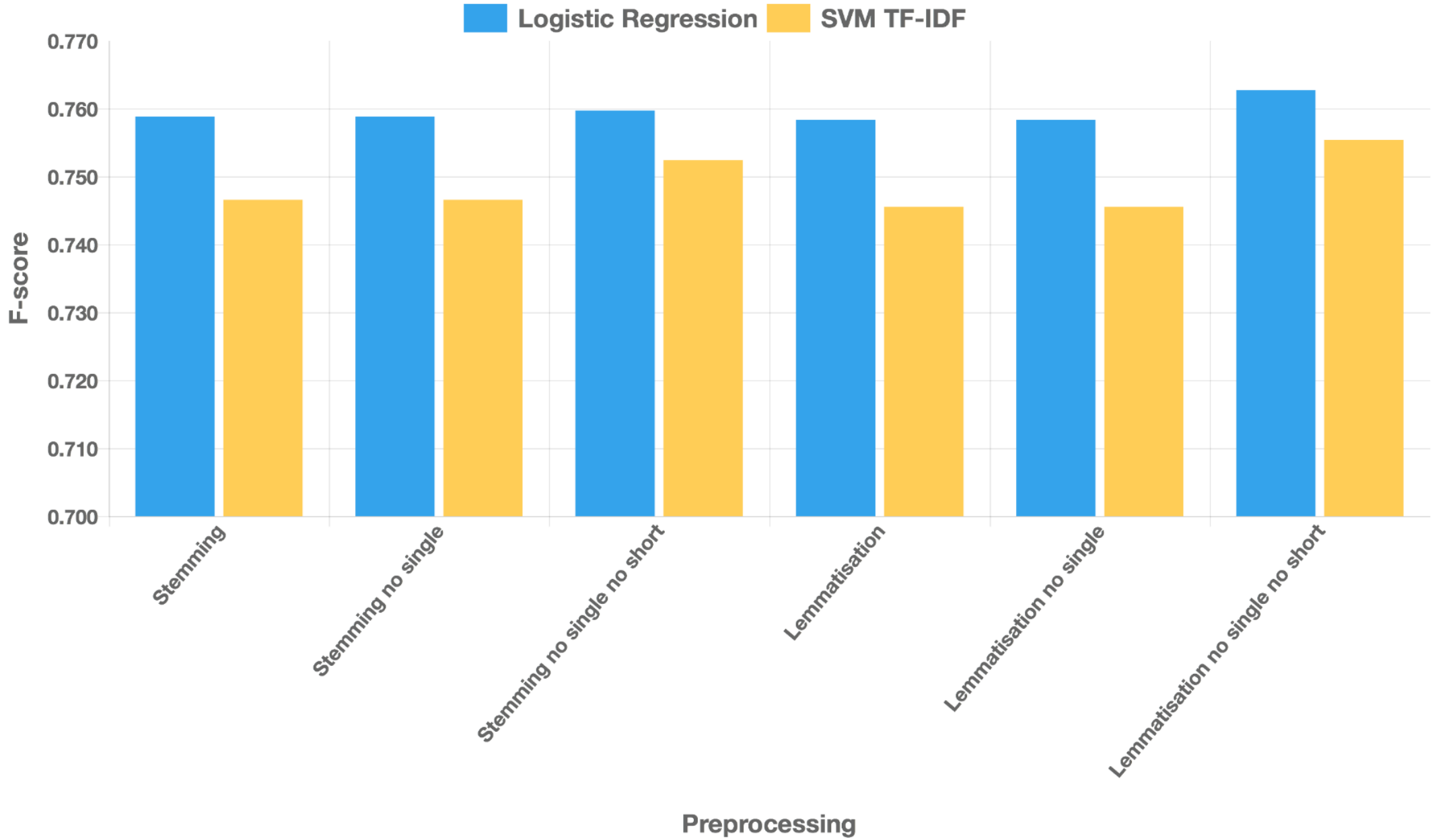
Основные параметры: ngram range, min_df

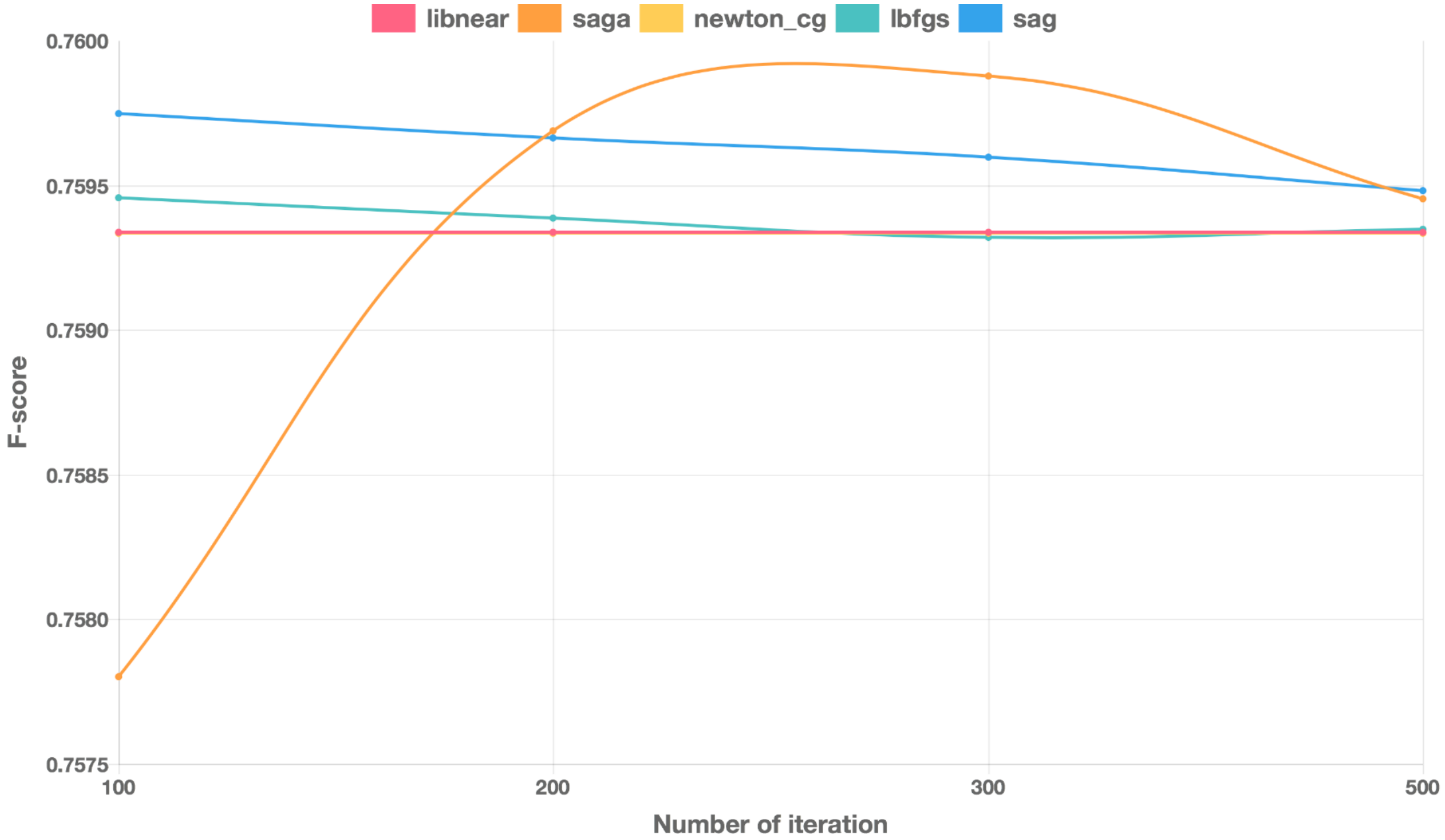


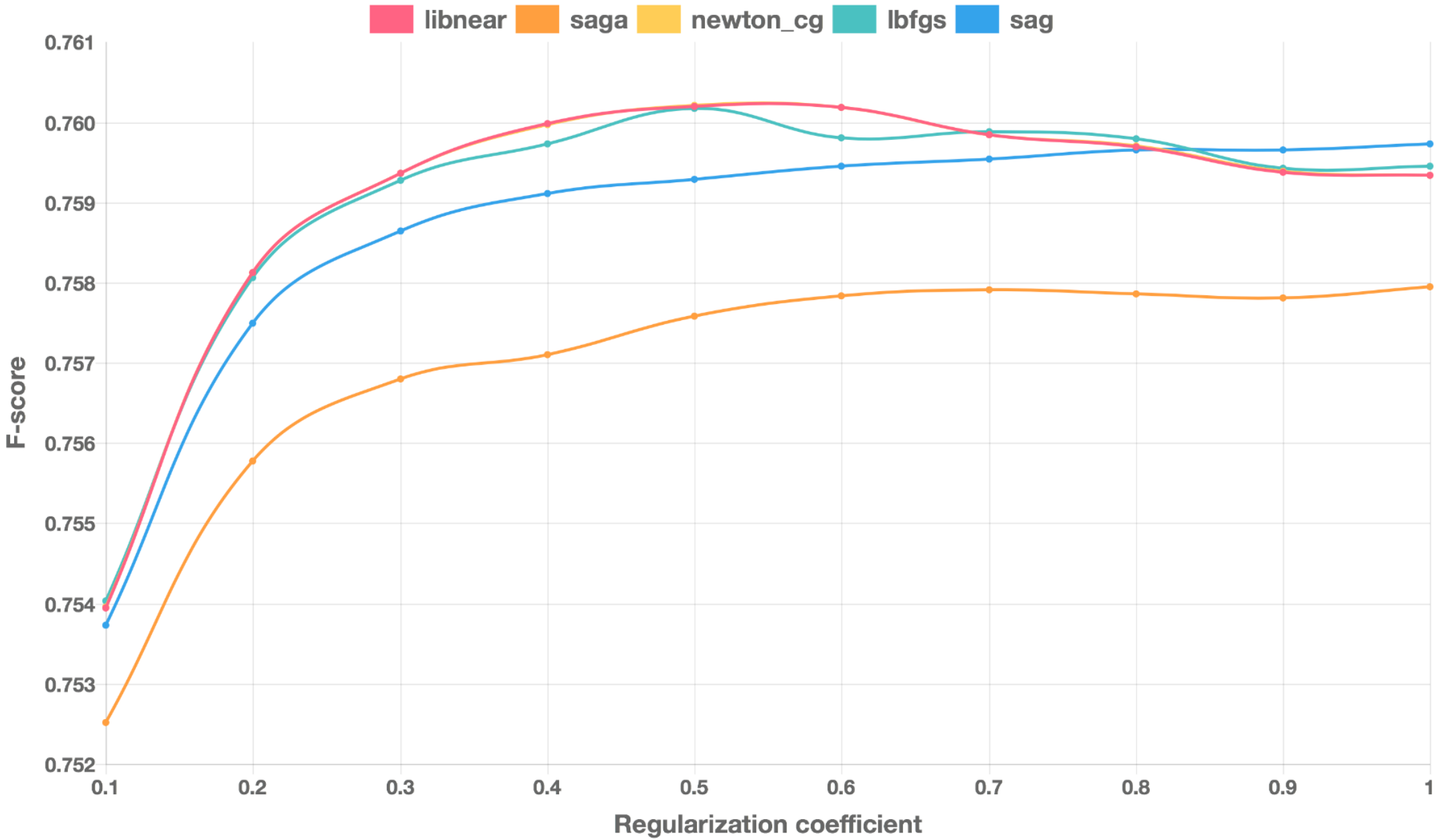
Улучшение результатов логистической регрессии

- Другой подход к обработке текста
  - Стемминг (NLTK) / Лемматизация (mystem)
  - Удаление слов из одной буквы
  - Удаление коротких твитов
- Параметры логистической регрессии
  - **penalty**
  - dual
  - **C**
  - fit_intercept
  - class_weight
  - **solver**
  - **max_iter**
  - multi_class
  - warm_start




Logistic Regression
SVM TF-IDF
0.770
0.760
0.750
0.740
0.730
0.720
0.710
0.700
F-score
Stemming
Stemming no single
Stemming no single no short
Lemmatisation
Lemmatisation no single
Lemmatisation no single no short
Preprocessing


libnear
saga
newton_cg
lbfgs
sag
0.7600
0.7595
0.7590
0.7585
0.7580
0.7575
F-score
100
200
300
500
Number of iteration

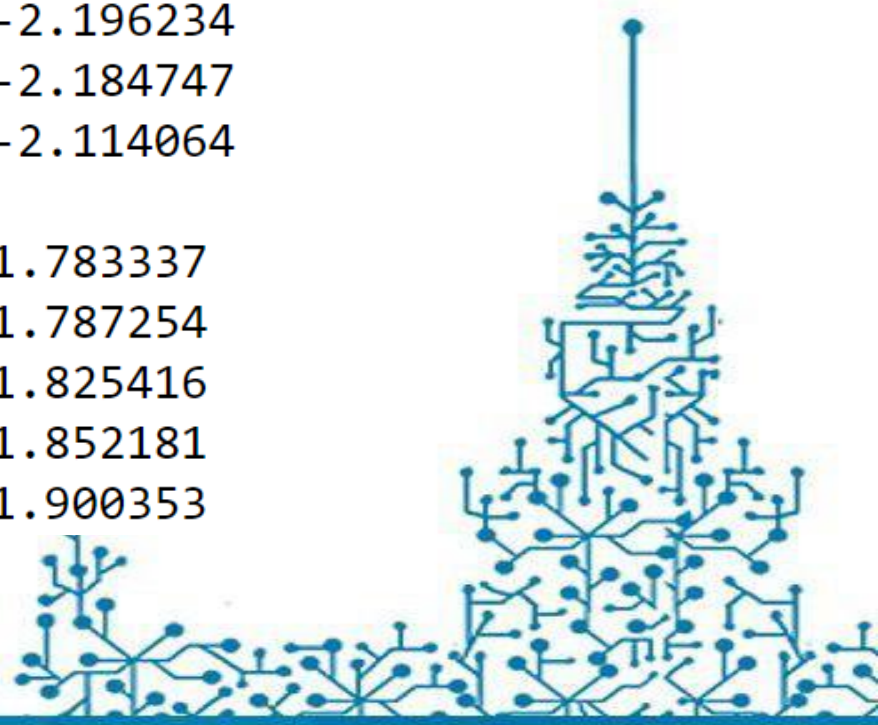

libnear
saga
newton_cg
lbfgs
sag
0.761
0.760
0.759
0.758
0.757
0.756
0.755
0.754
0.753
0.752
F-score
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Regularization coefficient

The best of the best of the best

- Обработка текста:
  - Лемматизация с отбрасыванием
  - Count Vectorizer
- Признаки
  - Word n-gram = (1,3)
  - Min-df = 1
- Модель
  - Логистическая регрессия
    - C = 0.5

Самые «тяжелые» слова модели

обидно	-2.593947
печально	-2.453825
грустно	-2.196234
расстраивать	-2.184747
печаль	-2.114064
...	
не забывать	1.783337
ахахахах	1.787254
приятно	1.825416
ахах	1.852181
не зря	1.900353



Сравнение результатов

Название работы	Имя авторов	Результат (F1 мера)
Классификация эмоциональной окраски сообщений в социальных сетях	Савинов Н.А	72.3%
Анализ тональности сообщений социальной сети twitter	Цветков А.Д.	76.2%
Sentiment Analysis of Twitter Data	Apoorv Agarwal, Boyi Xie, Ilia Vovsha	75.86%
Наша модель		76.4%
Человек в среднем		79%

Ссылки:

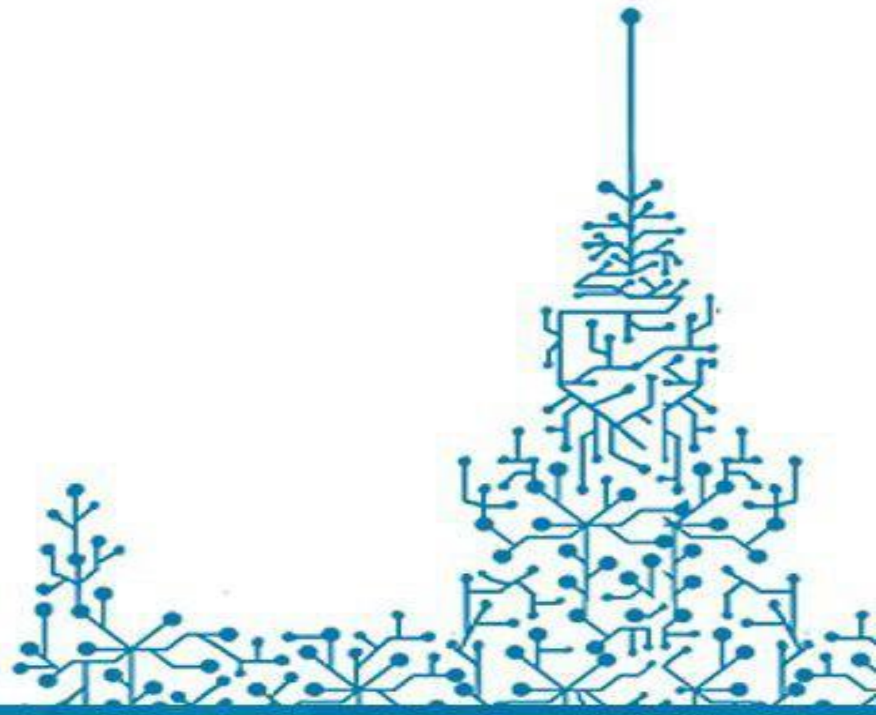
- http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.348.4054&rep=rep1&type=pdf
- http://www.csd.tsu.ru/sites/default/files/Хранилище/people/выпускники/2013/diplom_Svetkov.pdf
- http://www.machinelearning.ru/wiki/images/b/be/SavinovThesis2013.pdf
- http://mashable.com/2010/04/19/sentiment-analysis/#9tPwLHXSH5qz

Бизнес задачи

- Анализ отношения к продукту / компании / человеку
- Мониторинг отношения к событиям
- Реальные примеры:
  - Компания «Роснефть» объявила тендер на мониторинг СМИ и соцсетей по ключевому слову “Сечин”. Необходимо делить отзывы на три столбика - позитивные, отрицательные и нейтральные.
  - Исследования показывают, что между тональностью сообщений в социальных сетях, блогах и публикациях в медиа и положением дел на финансовых рынках действительно существует определенная связь (habrahabr)

Online

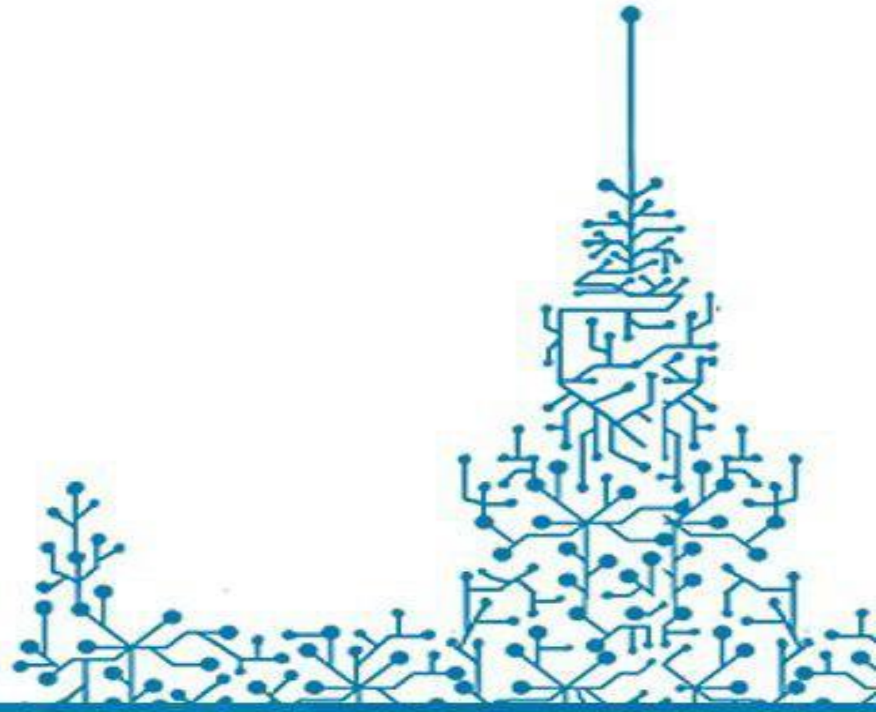
- Сервер: Django
- Обработка запросов: Javascript
- Представление: HTML 5, CSS
- Формат: json



Проект выполнили:

- Александр Смирнов
- Антон Лоскутов
- Иван Мажаров
- Максим Филин

Спасибо за внимание!



С НАСТУПАЮЩИМ

tutu