Sentiment Analysis using NLP to predict PE Ratio

Выполнила: Холодаева Е.В., 241 группа

Научный руководитель: Григорьев Д.А.

Санкт-Петербург, 2019

PE ratio

- **Отношение P / E (Price/Earnings** или **Цена/Прибыль)** - это отношение цены акции (акций) компании к ее прибыли на акцию. Коэффициент используется для оценки компаний и определения того, являются ли они переоцененными или недооцененными.

$$P/E = \frac{Share\ Price}{Earnings\ per\ Share}$$

Sentiment analysis

- **Анализ тональности текста (sentiment analysis)** – область компьютерной лингвистики, занимающаяся выделением из текстов эмоционально окрашенной лексики или эмоциональной оценки автора.

1. Подходы, основанные на правилах
2. Подходы, основанные на словарях
3. Машинное обучение с учителем
4. Машинное обучение без учителя

Цели и задачи работы

Целью курсовой работы является реализация методов анализа тональности текста и применение их для прогнозирования коэффициента PE

Указанная цель достигалась путем решения следующих задач:

1) Изучение основных методов анализа тональности текста

2) Их применение к данной задаче

3) Анализ полученных результатов

DataSet (78000 texts)

	A
1	ItemID,Sentiment,SentimentText
2	1,0, is so sad for my APL friend.....
3	2,0, I missed the New Moon trailer...
4	3,1, omg its already 7:30 :O
5	4,0, .. Omgaga. Im sooo im gunna CRy. I've been at this dentist since 11.. I was suposed 2 just get a crown put on (30mins)...
6	5,0, i think mi bf is cheating on me!!! T_T
7	6,0, or i just worry too much?
8	7,1, Juuuuuuuuuuuuuuuussssst Chillin!!
9	8,0, Sunny Again Work Tomorrow :-| TV Tonight
10	9,1, handed in my uniform today . i miss you already
11	10,1, hmmmm.... i wonder how she my number @-)
12	11,0, I must think about positive..
13	12,1, thanks to all the haters up in my face all day! 112-102

Тестовый набор/Обучающий набор = 1/100

https://www.kaggle.com/datasets

Очистка данных и предварительная обработка

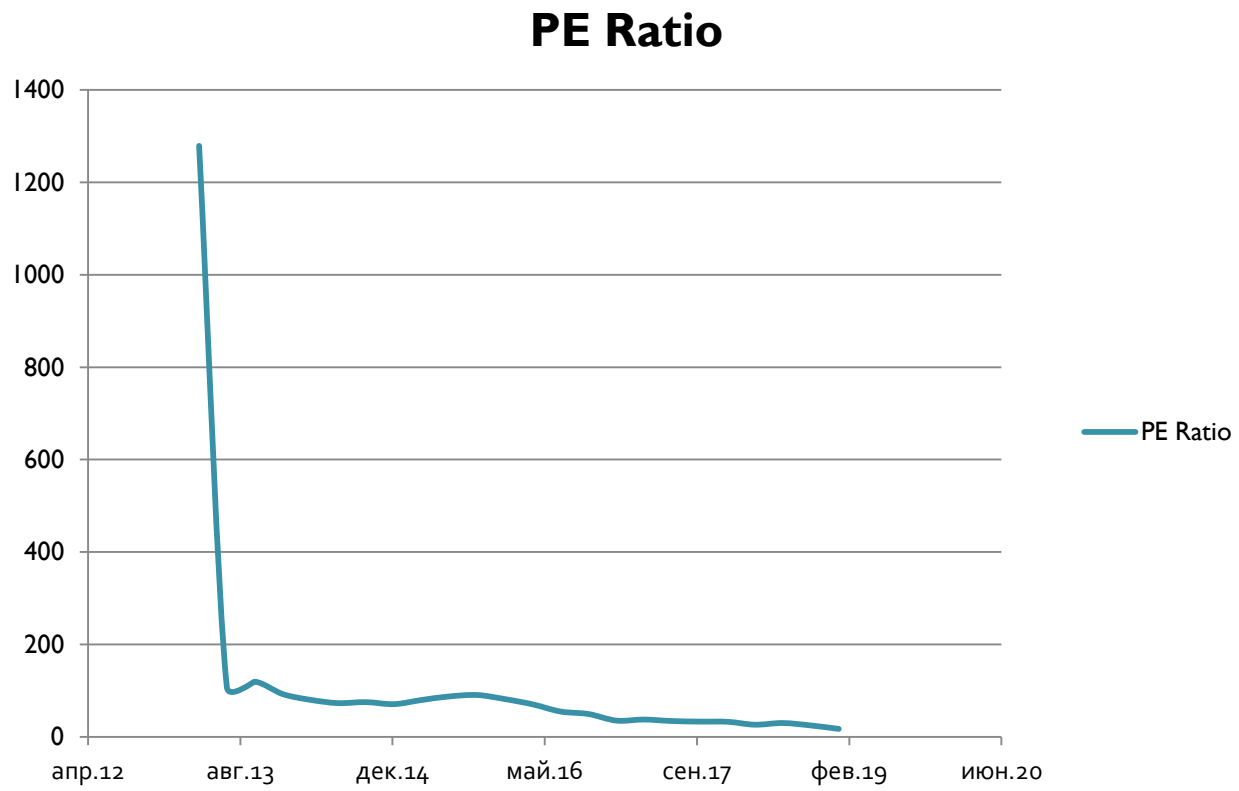
Удаление HTML-разметки

Удаление цифр и знаков препинания

Удаление стоп-слов(“the”,“a” и др.)

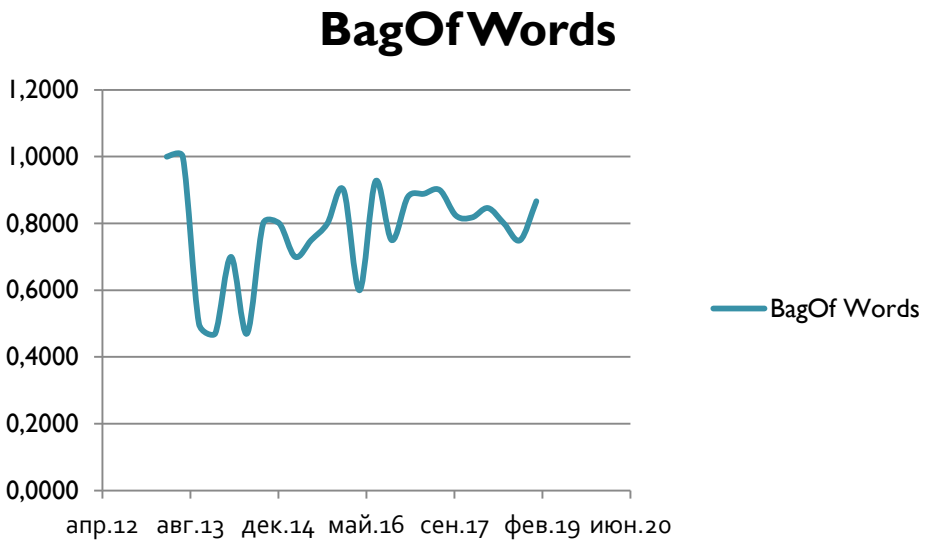
Стэмминг

Facebook PE Ratio 2013-2019

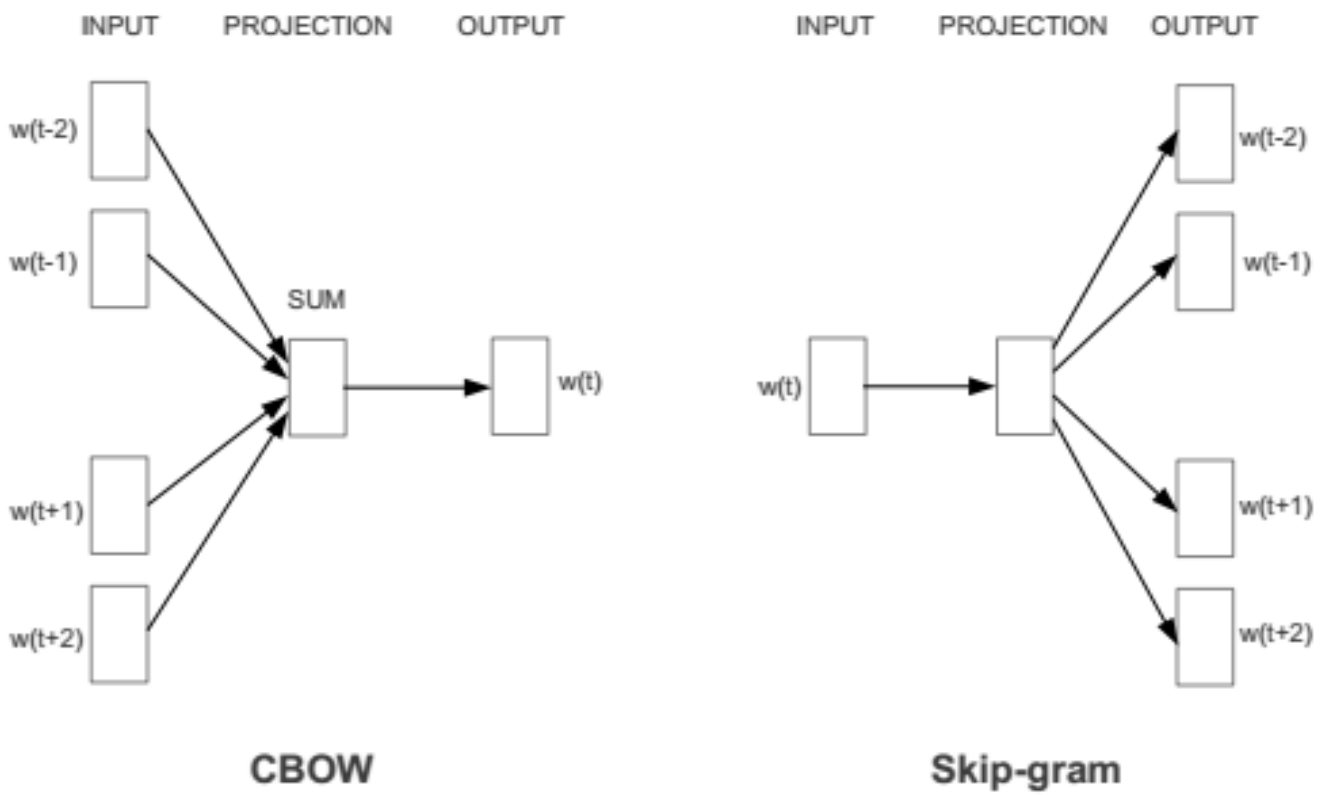


Bag Of Words

Date	Sentiment
апр.13	1,0000
июл.13	1,0000
окт.13	0,5000
янв.14	0,4700
апр.14	0,7000
июл.14	0,4700
окт.14	0,8000
янв.15	0,8000
апр.15	0,7000
июл.15	0,7500
окт.15	0,8000
янв.16	0,9000
апр.16	0,6000
июл.16	0,9265
окт.16	0,7500
янв.17	0,8800
апр.17	0,8889
июл.17	0,9000
окт.17	0,8235
янв.18	0,8182
апр.18	0,8462
июл.18	0,8000
окт.18	0,7500
янв.19	0,8667

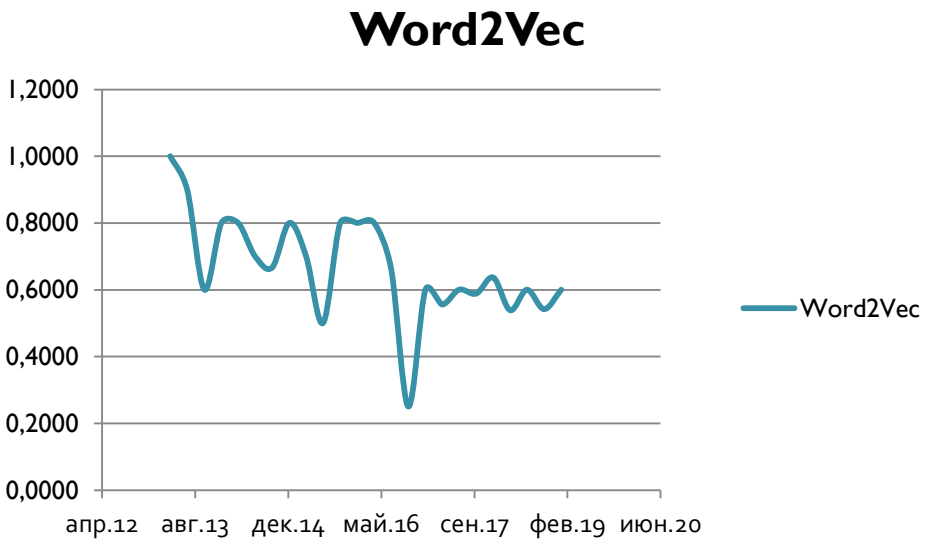


Word2Vec


INPUT
PROJECTION
OUTPUT
w(t-2)
w(t-1)
SUM
w(t)
w(t+1)
w(t+2)
CBOW
INPUT
PROJECTION
OUTPUT
w(t)
w(t-2)
w(t-1)
w(t+1)
w(t+2)
Skip-gram

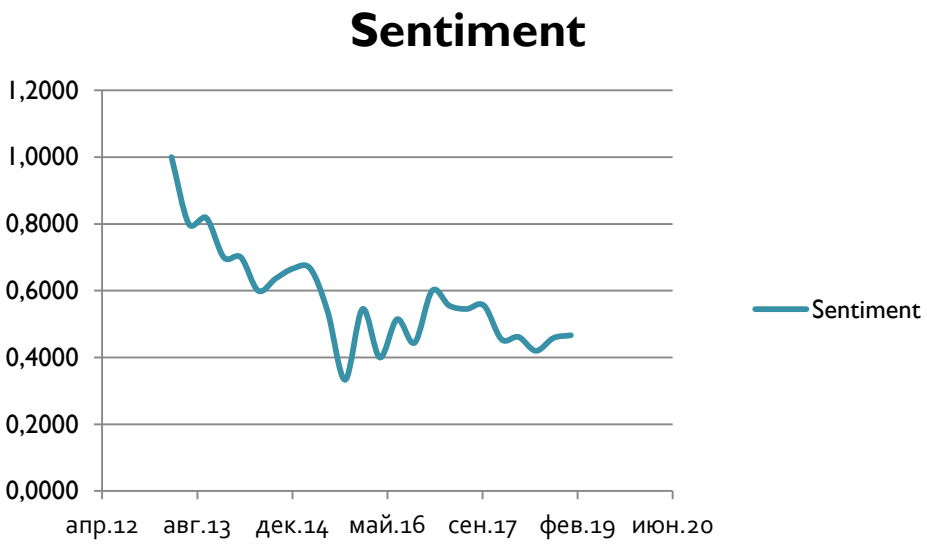
Word Vector

Date	Sentiment
апр.13	1,0000
июл.13	0,9000
окт.13	0,6000
янв.14	0,8000
апр.14	0,8000
июл.14	0,7000
окт.14	0,6667
янв.15	0,8000
апр.15	0,7000
июл.15	0,5000
окт.15	0,8000
янв.16	0,8000
апр.16	0,8000
июл.16	0,6618
окт.16	0,2500
янв.17	0,6000
апр.17	0,5556
июл.17	0,6000
окт.17	0,5882
янв.18	0,6364
апр.18	0,5385
июл.18	0,6000
окт.18	0,5417
янв.19	0,6000

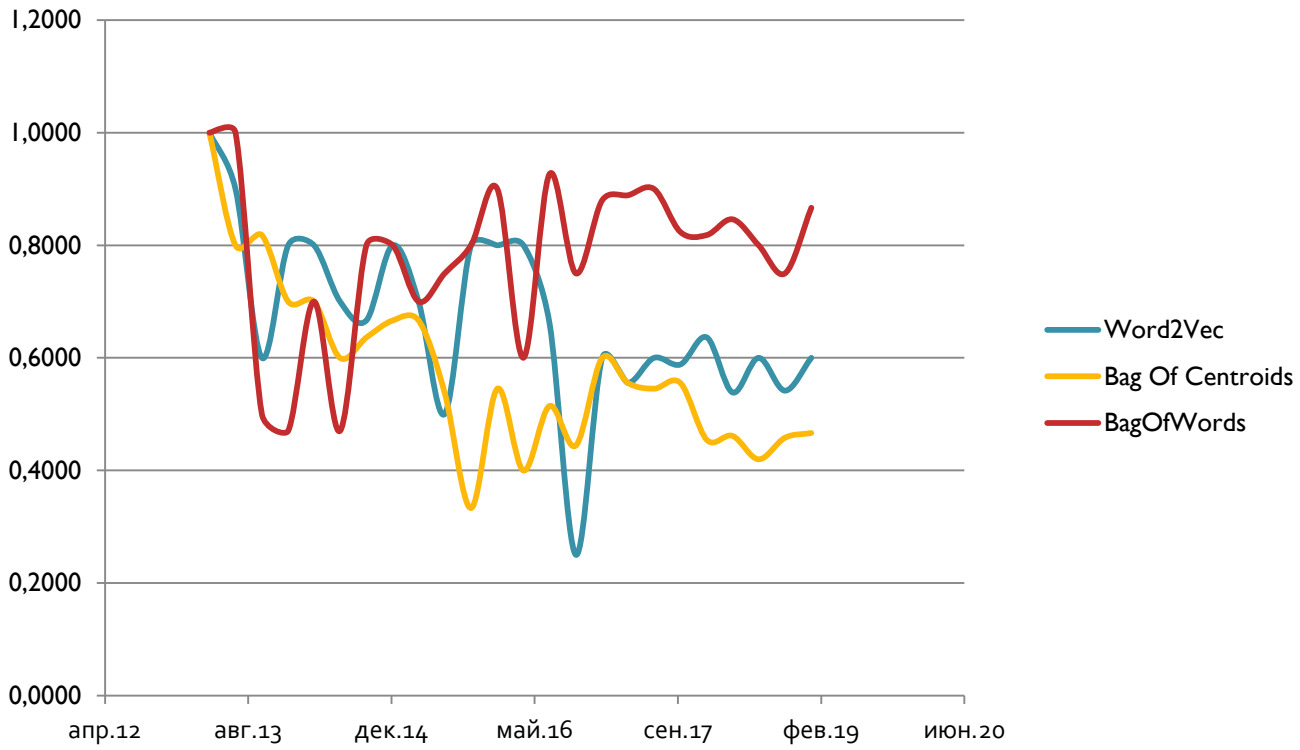


Bag Of Centroids

Date	Sentiment
апр.13	1,0000
июл.13	0,8000
окт.13	0,8182
янв.14	0,7000
апр.14	0,7000
июл.14	0,6000
окт.14	0,6364
янв.15	0,6667
апр.15	0,6667
июл.15	0,5385
окт.15	0,3333
янв.16	0,5455
апр.16	0,4000
июл.16	0,5147
окт.16	0,4444
янв.17	0,6000
апр.17	0,5556
июл.17	0,5455
окт.17	0,5556
янв.18	0,4545
апр.18	0,4615
июл.18	0,4200
окт.18	0,4583
янв.19	0,4667



Сравнение результатов



Корреляция графиков

Method	Correlation
Bag Of Words	0,268267
Word Vector	0,507743
Bag Of Cenroids	0,641367

Результаты

- Реализованы основные методы анализа тональности текста
- Прогнозирование индекса PE с помощью реализованных методов
- Произведен анализ результатов