Національний технічний університет України «КПІ ім. Ігоря Сікорського»

Факультет Інформатики та Обчислювальної Техніки

Кафедра Автоматизованих Систем Обробки Інформації та Управління

Лабораторна робота №2

з дисципліни «Комп'ютерна лінгвістика»

Тема:

«**АНАЛІЗ ТОНАЛЬНОСТІ ТЕКСТУ ЗА ДОПОМОГОЮ БАЄСІВА КЛАСИФІКАТОРА**»

Виконав:

студент гр. ІС-61

Іваницький Олег

Викладач: Фіногенов О.Д

Київ – 2019

## Мета роботи: вивчення методів класифікації при визначені тональності тексту.

**Варіант завдання:** день народження 20.03 (файл з даними у форматі csv прикріпллюється)

**Опис програми:**

Програма зчитує дані з csv файлу в pandas.DataFrame. Робиться копія даних. Перша копія очищається від зайвих знаків табуляції та пунктуації, друга, окрім того, лематизується (зведення слів до кореня та видалення стоп-слів\*). Після цього дані розбиваються на датасети: по 100, 120 та 140 екземплярів. Тестовий датасет містить 60 екземплярів даних. Тренувальні дані векторизуються, тобто кожному слову ставиться у відповідність кількість його вживання. Далі знаходиться число-показник приналежності повідомлення з тестової вибірки до певного класу. Це число розраховується як сума по усім словам повідомленням відношенням кількості вживання кожного слова в певному класі до загальної кількості слів у цьому класі. Прогнозований клас тестового повідомлення – клас, число-показник якого найбільший серед усіх класів. В кінці вираховується точність прогнозування і будуються графіки найпопулярніших слів для кожного з класів до і після лематизації.

\*Список стоп-слів, що використовуються у програмі (отримані з модулю stopwords бібліотеки nltk): и, в, во, не, что, он, на, я, с, со, как, а, то, все, она, так, его, но, да, ты, к, у, же, вы, за, бы, по, только, ее, мне, было, вот, от, меня, еще, нет, о, из, ему, теперь, когда, даже, ну, вдруг, ли, если, уже, или, ни, быть, был, него, до, вас, нибудь, опять, уж, вам, ведь, там, потом, себя, ничего, ей, может, они, тут, где, есть, надо, ней, для, мы, тебя, их, чем, была, сам, чтоб, без, будто, чего, раз, тоже, себе, под, будет, ж, тогда, кто, этот, того, потому, этого, какой, совсем, ним, здесь, этом, один, почти, мой, тем, чтобы, нее, сейчас, были, куда, зачем, всех, никогда, можно, при, наконец, два, об, другой, хоть, после, над, больше, тот, через, эти, нас, про, всего, них, какая, много, разве, три, эту, моя, впрочем, хорошо, свою, этой, перед, иногда, лучше, чуть, том, нельзя, такой, им, более, всегда, конечно, всю, между, это, з, й, що, та, із, але, цей, коли, як, чого, хоча, нам, якось, чи, це, від, їх, інших, ти, він, вона, воно, ми, ви, вони, якого, яких, яким, є, чому, чим, де, десь, собі, свій.

**Найбільш вживані слова:**

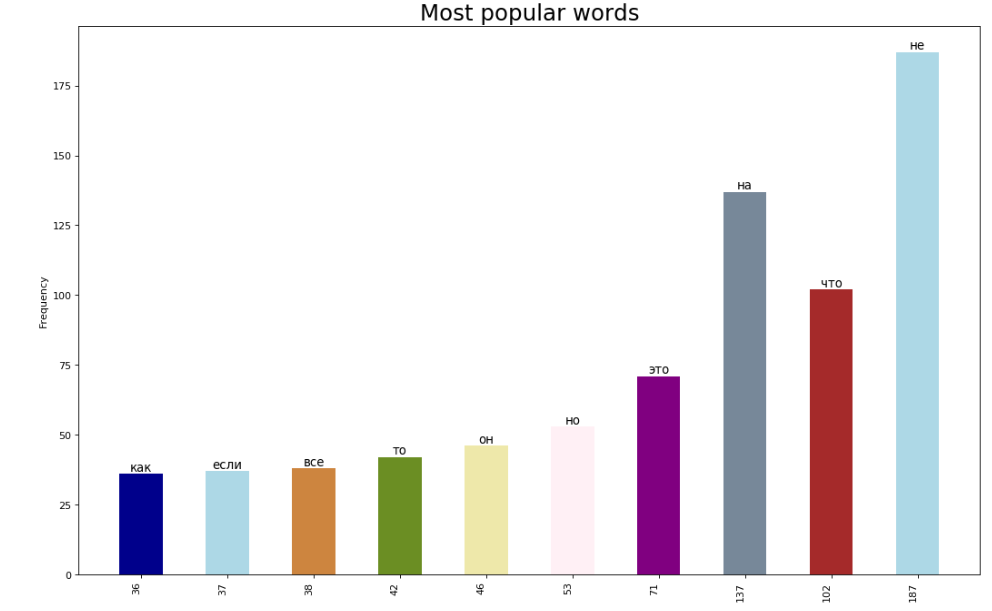


Рис.1 Найпопулярніші слова негативного класу до видалення стоп-слів

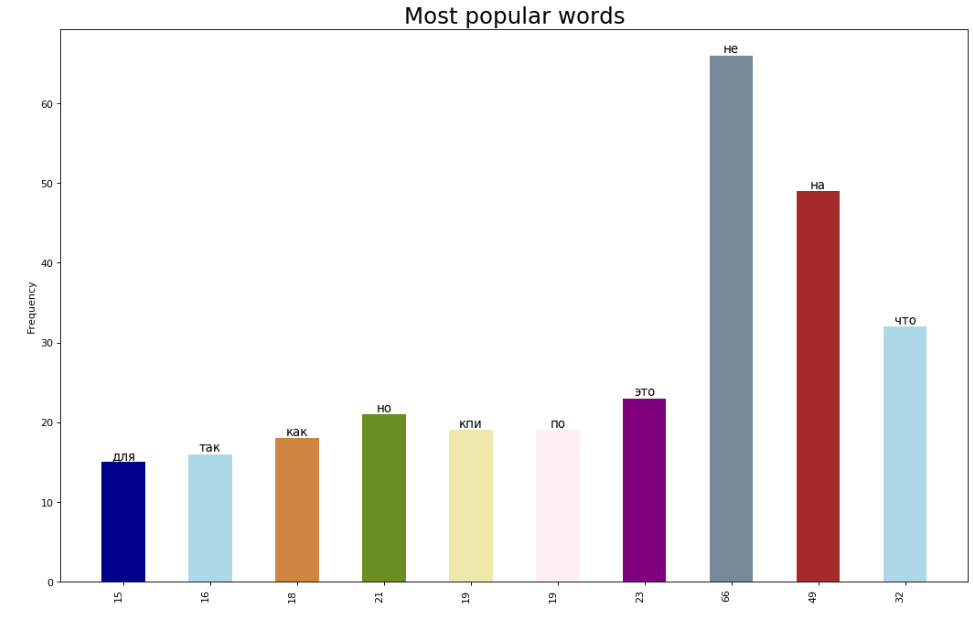


Рис.2 Найпопулярніші слова нейтрального класу до видалення стоп-слів

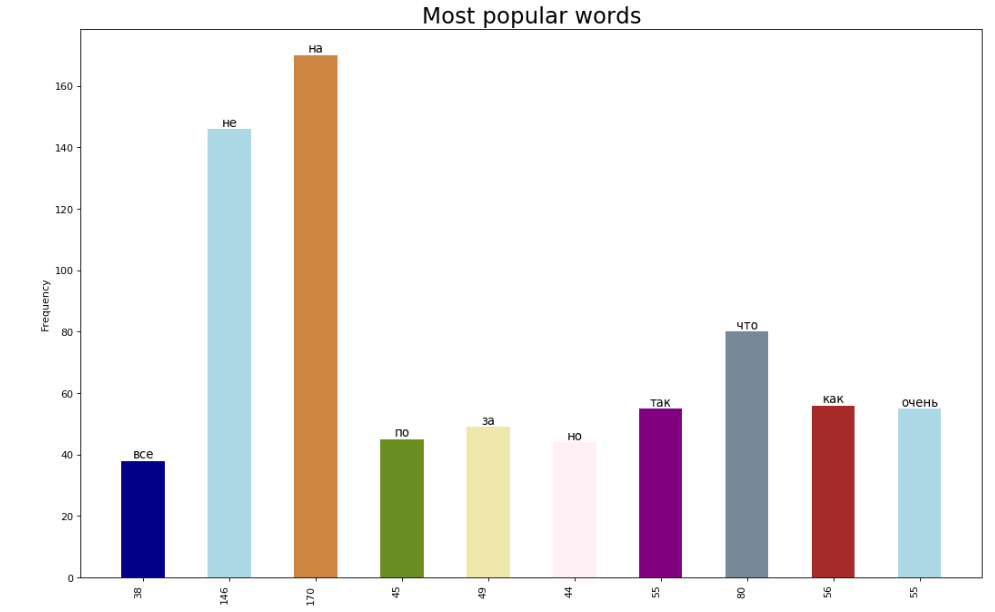


Рис.3 Найпопулярніші слова позитивного класу до видалення стоп-слів

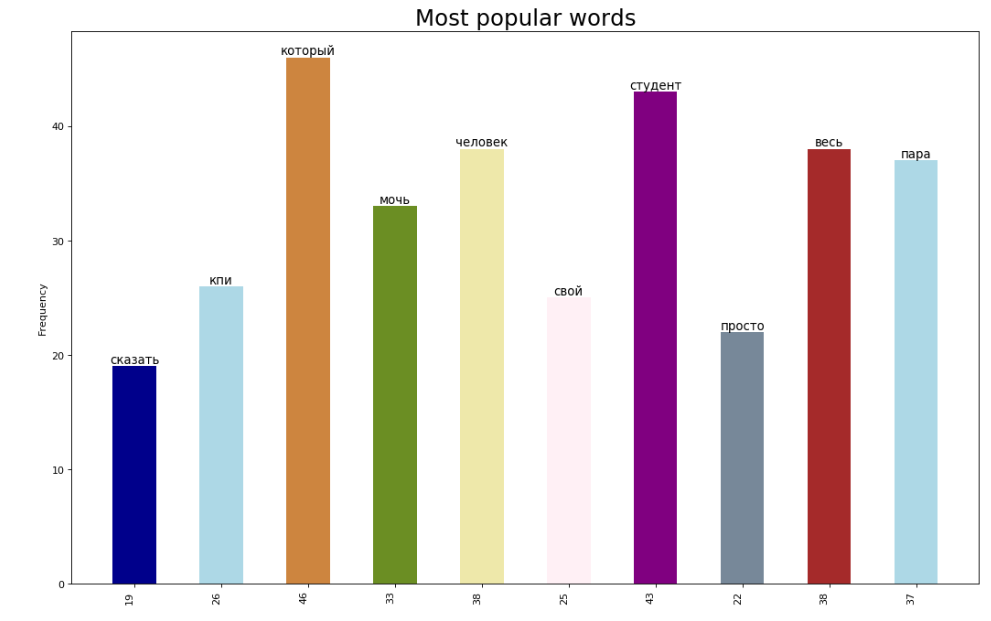


Рис.4 Найпопулярніші слова негативного класу після видалення стоп-слів

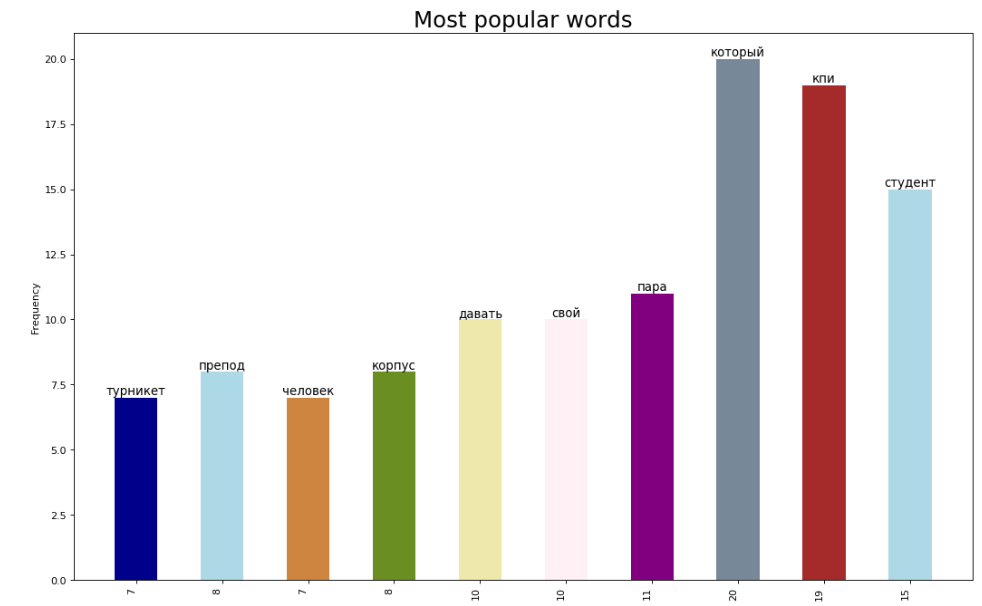


Рис.5 Найпопулярніші слова нейтрального класу після видалення стоп-слів

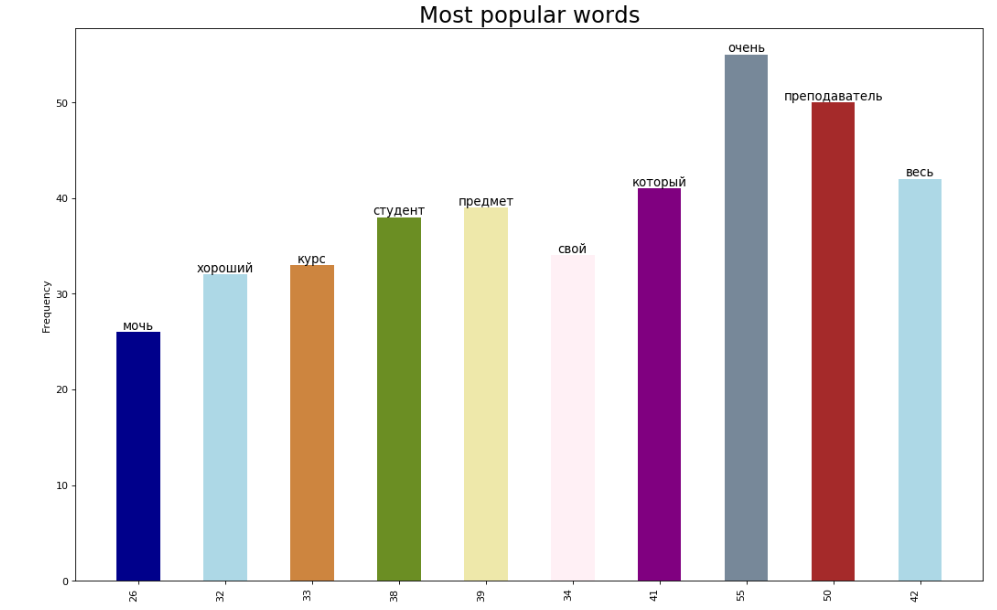


Рис.6 Найпопулярніші слова позитивного класу після видалення стоп-слів

**Точність класифікації:**

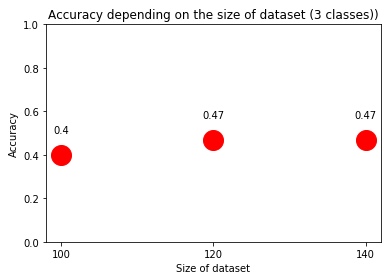


Рис.7 Залежність точності класифікації від розміру навчальної вибірки (3 класи)

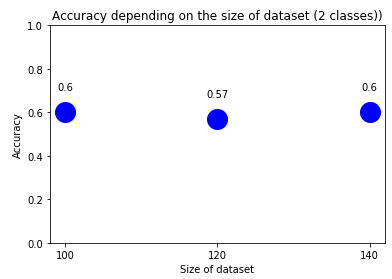


Рис.8 Залежність точності класифікації від розміру навчальної вибірки (2 класи)

**Висновки:** у ході лабораторної роботи було побудувано модель класифікації повідомлень, що визначає його настрій (позитивний, негативний або нейтральний). Протестувавши модель, навчену на навчальних вибірках різних розмірів, отримали найкращий результат – 60% точності. Цьому результату відповідає модель, навчена на вибірці зі 100 та 140 екземплярів даних, де є тільки два класи (позитивний та нейтральний). Таку невисоку точність класифікації можна виправдати неканонізованим текстом, що має різний формат, який дуже важко привести до уніфікованого вигляду. Окрім того, повідомлення написані двома мовами: українською та російською, а також суржиком, що негативно вплинуло на якість класифікації.