

## Лабораторна робота №6

### Аналіз настроїв

**Мета роботи:** Ознайомитись з вирішенням задачі аналізу настроїв.

#### Короткі теоретичні відомості

Текстові дані, незважаючи на те, що вони дуже неструктуровані, можна розділити на два основних типи документів. Фактичні документи, як правило, зображують певну форму тверджень або фактів без конкретних почуттів чи емоцій. Вони також відомі як об'єктивні документи. З іншого боку, суб'єктивні документи виражають почуття, настрої, емоції та думки.

Ключова ідея полягає в тому, щоб використовувати методи текстової аналітики, обробки природньої мови, машинного навчання та лінгвістики для вилучення важливої інформації або даних із неструктурованого тексту. Це, у свою чергу, може допомогти отримати якісні результати, як-от загальний настрої на позитивній, нейтральній або негативній шкалі, і кількісні результати, як-от полярність настроїв, пропорції суб'єктивності та об'єктивності. Полярність настроїв – це зазвичай числова оцінка, присвоєна позитивним і негативним аспектам текстового документа, яка базується на суб'єктивних параметрах, таких як конкретні слова та фрази, що виражають почуття та емоції.

Моделі аналізу настроїв без учителя використовують добре підібрані бази знань, онтології, лексикони та бази даних, які містять детальну інформацію, що стосується суб'єктивних слів, фраз, включаючи почуття, настрої, полярність, об'єктивність, суб'єктивність тощо. Лексикони містять список слів, пов'язаних із позитивним і негативним настроєм, полярністю (величиною негативного чи позитивного результату), тегами частин мови, класифікаторами суб'єктивності (сильний, слабкий, нейтральний), настроєм, модальністю тощо.

Використаємо лексикон TextBlob та набір даних з відгуками на фільми.

```
for review, sentiment in zip(test_reviews[sample_review_ids],
test_sentiments[sample_review_ids]):
```

```
    print('REVIEW:', review)
```

```
    print('Actual Sentiment:', sentiment)
```

```
    print('Predicted Sentiment polarity:', textblob.TextBlob(review).
sentiment.polarity)
```

```
    sentiment_polarity = [textblob.TextBlob(review).sentiment.polarity for
review in test_reviews]
```

```
    predicted_sentiments = ['positive' if score >= 0.1 else 'negative' for score in
sentiment_polarity]
```

Ще один спосіб побудувати модель для розуміння текстового вмісту та прогнозування настрою текстових відгуків — це використовувати машинне навчання з учителем. Точніше, використовувати класифікаційні моделі для вирішення цієї проблеми.

```
norm_train_reviews = tn.normalize_corpus(train_reviews)
```

```

norm_test_reviews = tn.normalize_corpus(test_reviews)
cv = CountVectorizer(binary=False, min_df=0.0, max_df=1.0,
ngram_range=(1,2))
cv_train_features = cv.fit_transform(norm_train_reviews)
tv = TfidfVectorizer(use_idf=True, min_df=0.0, max_df=1.0,
ngram_range=(1,2), sublinear_tf=True)
tv_train_features = tv.fit_transform(norm_train_reviews)
cv_test_features = cv.transform(norm_test_reviews)
tv_test_features = tv.transform(norm_test_reviews)
from sklearn.linear_model import SVC, LogisticRegression
lr = LogisticRegression(penalty='l2', max_iter=100, C=1)
svm = SVC(loss='hinge', max_iter=100)

```

### Завдання до лабораторної роботи

Створити програму, яка:

1. Зчитує заданий набір даних, виконує попередню обробку, розбиває дані на навчальні та тестові. Виконує аналіз настроїв за допомогою алгоритмів класифікації (наприклад, логістичної регресії, опорних векторів і т.д.).
2. Розрахувати матрицю невідповідностей, провести оцінку точності моделі.
3. Використати один з готових лексиконів, наприклад Textblob, для аналізу оцінки настроїв. Також розрахувати матрицю невідповідностей, провести оцінку точності моделі.
4. Обрати три випадкові записи та вивести результати оцінки їх настрою за пунктами 1 і 3.

Варіант 1.

Файл twitter1.csv. Використати логістичну регресію

Варіант 2.

Файл Reddit.csv. -1 – негативний коментар, 0 – нейтральний, 1 – позитивний. Використати опорні вектори

Варіант 3.

Файл movie1.csv. 0 – негативний коментар, 1 – позитивний. Використати логістичну регресію

Варіант 4.

Файл twitter2.csv. -1 – негативний коментар, 0 – нейтральний, 1 – позитивний. Використати наївний байєсів класифікатор

Варіант 5.

Файл movie2.csv. 0 – негативний коментар, 1 – позитивний. Використати опорні вектори

Варіант 6.

Файл twitter1.csv. Використати опорні вектори

Варіант 7.

Файл Reddit.csv. -1 – негативний коментар, 0 – нейтральний, 1 – позитивний. Використати наївний байєсів класифікатор

Варіант 8.

Файл movie1.csv. 0 – негативний коментар, 1 – позитивний. Використати опорні вектори

Варіант 9.

Файл twitter2.csv. -1 – негативний коментар, 0 – нейтральний, 1 – позитивний. Використати логістичну регресію

Варіант 10.

Файл movie2.csv. 0 – негативний коментар, 1 – позитивний. Використати наївний байєсів класифікатор

Варіант 11.

Файл twitter1.csv. Використати наївний байєсів класифікатор

Варіант 12.

Файл Reddit.csv. -1 – негативний коментар, 0 – нейтральний, 1 – позитивний. Використати логістичну регресію

Варіант 13.

Файл movie1.csv. 0 – негативний коментар, 1 – позитивний. Використати наївний байєсів класифікатор

Варіант 14.

Файл twitter2.csv. -1 – негативний коментар, 0 – нейтральний, 1 – позитивний. Використати наївний байєсів класифікатор

Варіант 15.

Файл movie2.csv. 0 – негативний коментар, 1 – позитивний. Використати логістичну регресію

Оформити звіт. Звіт повинен містити:

- титульний лист;
- код програми;
- результати виконання коду;

Продемонструвати роботу програми та відповісти на питання стосовно теоретичних відомостей та роботи програми.