МОДУЛЬ АНАЛІЗУ ТА ЗБОРУ НЕГАТИВНИХ ВІДГУКІВ

СТУДЕНТКА ГРУПИ КА-35

ВЕРИГА КАТЕРИНА

АКТУАЛЬНІСТЬ

- Отримання рекомендацій щодо покупки
- Фільтрація рецензій
- Усереднена оцінка деякого предмета

ОБ'ЄКТ ТА ПРЕДМЕТ ДОСЛІДЖЕННЯ

- Об'єкт: методи класифікації документів
- Предмет: модуль, призначений для збору та аналізу текстових документів

ПОСТАНОВКА ЗАДАЧІ

- Проаналізувати існуючі рішення аналізу тональності тексту та їх переваги й недоліки
- Проаналізувати алгоритми класифікації на основі машинного навчання
- Обрати оптимальний алгоритм чи композицію
- Реалізувати програмний продукт згідно з темою

АЛГОРИТМ РОБОТИ СИСТЕМИ

- Первинна підготовка та навчання
  - Зібрати та розмітити навчальні тексти
  - Здійснити нормалізацію отриманих документів(стемінг)
  - Привести до векторного вигляду
  - Розбити вибірку на навчальну та валідаційну частини
  - Подати дані на вхід класифікатора
  - Визначити оптимальний класифікатор
- Подальше використання
  - По запиту користувача виконати аналіз тональності знайдених відгуків

ЗБІР ІНФОРМАЦІЇ

- Достатній об'єм початкової навчальної вибірки
- Стемінг
- Виключення стоп слів

ВИБІР ОЗНАК

- **На основі частоти виникнення слова.**

виключаємо з простору ознак все слова, які зустрічалися в навчальній вибірці менш м раз. Число m підбирається експериментально.

- **На основі взаємної інформації (mutual information)**

Значення взаємної інформації показує взаємозв'язок ознаки і класу. Для кожного класу береться м ознак з найбільшими значеннями.

НАЇВНИЙ БАЙЄСІВСЬКИЙ КЛАСИФІКАТОР

- ймовірнісний класифікатор, що використовує теорему Баєса для визначення ймовірності приналежності спостереження (елемента вибірки) до одного з класів при припущенні (наївному) незалежності змінних.

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

- оперує виходячи з таблиці ваг, які він отримує на етапі навчання з частотності класифікаційних ознак в межах класів, - по одному параметру на кожну пару класифікаційна ознака + клас

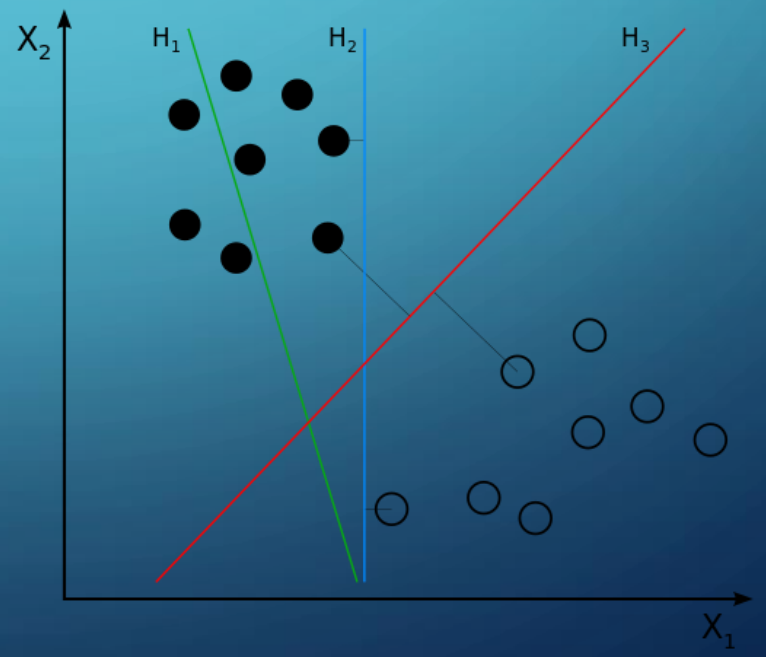
МЕТОД МАКСИМАЛЬНОЇ ЕНТРОПІЇ

- він ґрунтується на взаємозв'язку ентропії імовірнісного розподілу і його рівномірності;
- він використовує ітеративні алгоритми підбору параметрів моделі, які складніше в реалізації і більш ресурсомісткі.

$$p(c \mid d, \lambda) = \frac{\exp \sum_i^{n \times k} \lambda_i f_i(c, d)}{\sum_{\tilde{c} \in C} \exp \sum_i^{n \times k} \lambda_i f_i(\tilde{c}, d)}$$

SVM (МЕТОД ОПОРНИХ ВЕКТОРІВ)

- вихідні вектори переводяться в простір більш високої розмірності і шукаються роздільні гіперплощини з максимальним проміжком між ними.
- гарантує єдиність розв’язку



КОМПОЗИЦІЯ

Для підвищення точності класифікації розглядають ансамблі алгоримів.

- Беггінг
- Бустинг - мета-алгоритм машинного навчання ансамблів класифікаторів, що дозволяє підвищити точність класифікації,
  - є адаптивним алгоритмом - кожен наступний класифікатор будується по об'єктах, невірно класифікованих попередніми класифікаторами.
  - Алгоритм бустінгу застосовується для лінійних класифікаторів.

РЕЗУЛЬТАТ ДОСЛІДЖЕНЬ

	Precision	Recall	F-measure
CV NBmultinomial	0.85	0.84	0.84
CV NB_Bernoulli	0.83	0.81	0.84
CV MaxEnt	0.87	0.87	0.87
CV SVM	0.86	0.86	0.86
TF-IDF NBmultinomial	0.82	0.78	0.75
TF-IDF NB_Bernoulli	0.83	0.81	0.80
TF-IDF MaxEnt	0.86	0.86	0.86
TF-IDF SVM	0.86	0.86	0.86

АНАЛІЗ ТОЧНОСТІ

- Precision - частка результатів, яка дійсно належить даному класу
- Recall - відсоток знайдених результатів від їх загального числа
- Accuracy – правдоподібність
- F-measure - гармонічне середнє точності і повноти

ВИСНОВКИ

- Було досліджено існуючі рішення аналізу тональності тексту
- Обрано оптимальні алгоритми для текстів вхідної вибірки
- Розроблений програмний продукт, що використовує знайдені алгоритми класифікації на базі машинного навчання
- Модуль може
  - визначати поданий текст заданої тематики як негативний чи позитивний
  - Фільтрувати вхідні документи за вказаним критерієм(наприклад, негативні)

ШЛЯХИ ПОДАЛЬШОГО РОЗВИТКУ

- Збільшення навчальної вибірки для підвищення точності
- Навчити алгоритм для різних тематик

ДЯКУЮ ЗА УВАГУ!