Курсов проект по Подходи за обработка на естествен език

Изготвил Румен Чолаков

Анализ на настроенията в туитове

Задача

Извличане на туитове за определена тематика и определяне на изразеното в тях настроение.

> Работата по извличането на туитовете е задача за предмета Извличане на информация. В текущия момент този модул от проекта е неактивен и за обучение и тестване се иползват предварително класифицирани набори от данни, които могат да бъдат намерени на този адрес

Реализация

Елементи

Проектът е реализаиран на Python, а основните библиотеки, които са използвани са NLTK и scikit-learn.

- NLTK - предоставя инструментите нужни за предварителна обработка на данните, за да бъдат приведени във вид подходящ за употреба. Използвани комоненти
  - stopwords - предоставя колекция от стоп думи за английския език, които трябва да бъдат пермахнати от данните.
  - WordNetLemmatizer - предоставя възможност за лематизация на срещнатите думи използвайки WordNet
- scikit-learn - предоставя удобни за употреба модели за МС, който могат бъдат обучавани да класифицират добре форматирани данни. Използвани комоненти
  - CountVectorizer - трансформира данните в матрица показваща броя на срещания на отделните думи в данните
  - TfidfTransformer - отеглява стаойностите в матрицата използвайки честотата на срещане на отделните думи както в единични файлове(туитове), така и в набора от данни като цяло
  - MultinomialNB - реализация на наивен бейсов класификатор, подходяща за класифициране на текст преминал нужната предварителна обработка

Процес на работа

- Необработените данни (позитивни и негативни примери) се прочитат и обработват. Всеки от туитовете бива изчистван от препинателни занаци, именан на потребители и стоп думи, след което останалие в него думи се прекарват през WordNetLemmatizer, който ги заменя с подходяща основна форма на съответаната дума и накрая се добавя класът на обработвания пример. Така получените данни се записват в csv файл, който се ползва за вход на следващитете етапи.
- Обработените данни се подават на CountVectorizer и TfidfTransformer, който кодират думите до матрици от стойности подходящи за обработка от MultinomialNB
- MultinomialNB се обучава върху 75% от наличните данни и се тества върху останалите 25%

- Накрая се извеждат статистики за точността на MultinomialNB върху наличните данни

Кодът на проекта може да бъде намерен тук