**СУ „Св. Климент Охридски“**

Факултет по математика и информатика

**КУРСОВ ПРОЕКТ**

**НА ТЕМА :**

Оценяване на изречение за положителен или отрицателен смисъл

Петър Динев 80427

Мартин Иванов 80401

# Цел

Целата на нашия проект е да демонстрита разпознаване на положителния или отрецателни смисъл. Той пресмята силата на позитевни и негавитни емоции в кратки текстове дори и за неформален език или жаргон. Проекта ще излопзва няколко речника със оценени думи, всяка от които е асоцииирана със тегло което е нейната смислова тежест. В добавка този подход ползваме и някои нестандартни граматически правила. Използван е описан в Web пространството метод за класификация на мнение.

# Същност

Да сравни резултатите от осъществения алгоритъм с вече съществущ програмен продукт – SentiStrength. При подадено изречение приложението търси думите в изричението в базата с вече оценение думи, като се вземат предвид отрицания, емотикони, словосъчетания и дори има възможност за жаргон. Приложението намира цялостния максимлен положителен и отрицателен смисъл оценен със в съотвествие оценките на думите. Ако думата не се намира в базата се обръща към twitter. Намират публични постове в който се съдържа съответната дума ( подобрение ще е разпознаването на фрази ). На базата на няколкостотин поста ( задават се като параметър ) се извлича по-честия контекст на дадена дума в социалната мрежа Twitter. С помоща на SentiStrength се оценяват контекстите от посочените постове и се добавят към крайния резултат на изречението което се използва следната формула positive / ( positive + negative ) \* 10 – 5 за да се получи краен резултат който да е м/у -5 и 5 и да може да се приложи адекватно към оценката на изречението до момента.

Описание

Основните ресурси в алгоритъма са :

-речник от вече оценени 890 думи със стоности от -5 до 5

- речник със подсилващи думи или заслабващи смисъла думи;

- кратък списък с идиоми и често срещани фрази

- Речник от негативни думи обръщащи смисъла; списък емо-икони отново с оценка

-извличане на момента на непознатите думи от Web

-класифициране според контекстта им в Twitter

Резултатите на класификатора ни са бъдата 3 категории – от слабо до силно позитивни, негативни и неутрални. За да класифицираме изреченията в тези класове използваме API на продукта SentiStrength, при който основните групи са отделни думи, комбинации на прилагателни, отрицания, съюзи и предефинирани емот-икони.

Архитектура

Приложението се състои от два модула написани съответно на програмни езици: python и java. Модула на Java се грижи за извличането на информация от twitter както и оценяването ѝ като положителна или отрицателна(чрез SentiStrength) и оценява цяло изречение. Модула написан на Python изпраща заявка към Java модула като праща дума и получава оценката ѝ, зарежда речниците за да придобие представа кои от тях се включват в оценката.

Нефункционални характиристки

За осъществяването на проекта се ползва следните технологии – Python 2.7, sqlite3. Python е език от високо ниво, интерпретативен и подходящ точно за целите на проекта. sqlite3 е лесно достъпна технология, допълваща нуждите на заданието. За да разглеждаме тестовите данни които влизат в базата данни сме използвали плъгин на Mozzila Firefox SqLightManager. Използвани са и библияотеките twitter4J, jtwitter, sentistrength-2.2.