АЛГОРИТАМ ЗА ЕКСТРАКЦИЈА НА СТРУЧНИ ТЕРМИНИ ОД ТЕКСТ

Проф. д-р Ванчо Чабуковски
cabukv@hotmail.com

Природно-математички факултет, УКИМ - Скопје

Семинар “Алгоритми и примена” 10.12.2016

Вовед

- Проект: “Развој на систем за издвојување (екстракција) на стручни термини од текстови на македонски и албански јазик”, ПМФ, УКИМ-Скопје и ФПМН, ДУТ-Тетово
- Детектирање на проблемот
- Потреба – терминолошки речници
- Решение – компоненти на системот
- Активности

Вовед (прод.)

- Податочна анализа
- Издвојување на сентимент
- Масивни бази на податоци
- Пресметување во облак

Вовед (прод.)

- Формирање на корпуси на термини врз основа на честота на појавување
- Адаптибилност при селекцијата (можност за учење)
- Прочистување на базите (автоматско/мануелно)
- Спарување на термините (автоматски - на база на исти преведен текст, мануелно спојување на термините)

Алгоритам / Хевристика

Стручни текстови

1. Селекција преку препознавање на стручен текст од одредена област
2. Парсирање / пресметување на фреквенции / сортирање

Стручни термини
подередени според фреквенција на појавување од max до min

Праг: f > 5

Слободни термини
подередени според фреквенција на појавување од max до min

Праг: f < 3

Слободни текстови

1. Селекција преку препознавање на слободен текст
2. Парсирање / пресметување на фреквенции / сортирање

Алгоритам / Хевристика (стручни 60%, слободни 40%)

Подобрување на алгоритмот

- Зголемување на валидноста
- Зголемување на релевантноста
- Автоматско формирање на праговите
- Скратување на времето на учење
- Лесна надградливост

Подобрување на алгоритмот

- Релативна фреквенција:

$$RF(W) = F(W) / S(TXT)$$

RF(W) – релативна вреквенција на зборот W
F(W) – фреквенција на збор во одреден текст
S(TXT) – големина на текстот во број на зборови

- Формирање на праговите:

Termhood (T(W)) – степен на повразност на еден збор со одредена област т.е. проценка на јачината колку еден збор е стручен термин:

$$T(W) = RFS(W) / RFG(W)$$

RFS(W) – релативна фреквенција на W како стручен термин
RFG(W) – релативна фреквенција на W како слободен термин

$$TR = ((HT+LT)/2)+LT$$

HT – максималниот termhood во базата на податоци
LT – минималниот termhood во базата на податоци
TR – праг на значајност за екстракција на стручен термин

Подобрување на алгоритмот

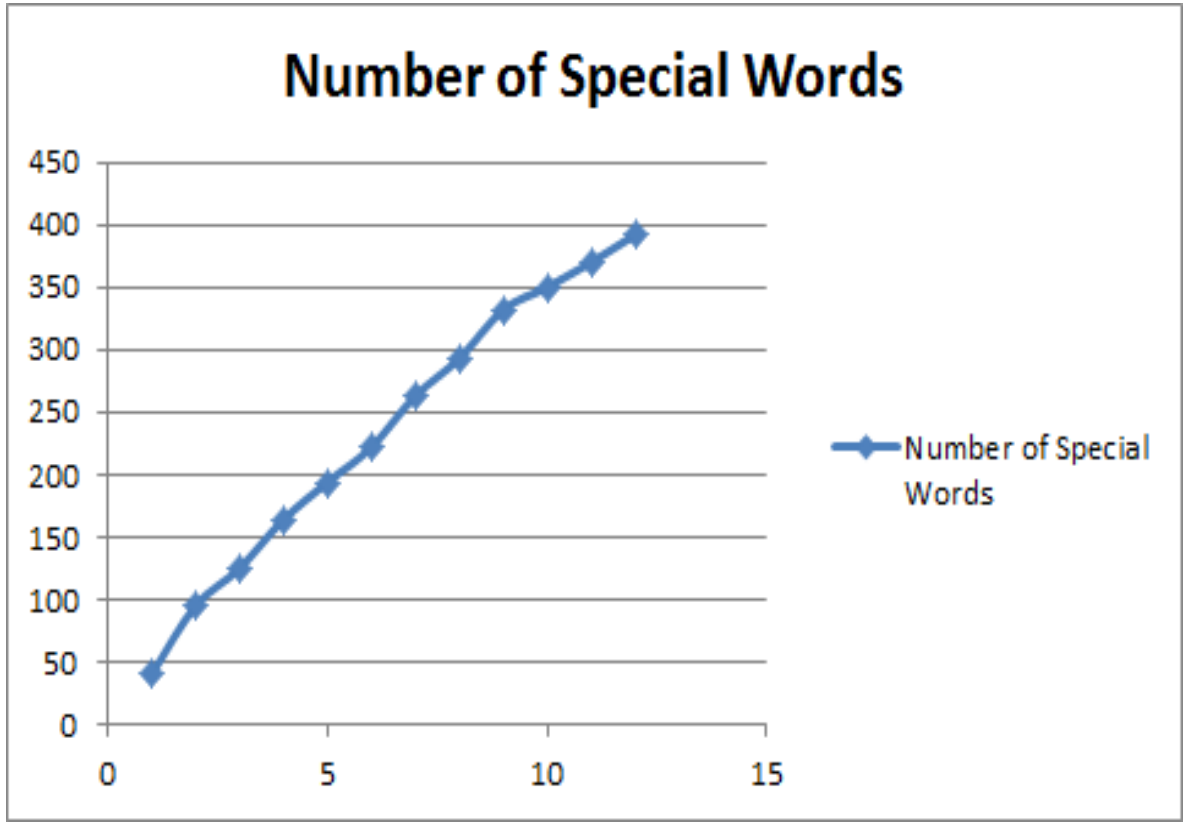
- Ако **T(W) >= TR**, W е стручен термин
- Ако **T(W) < TR**, W не стручен термин
- TR се прсметува при секое внесување на нов текст
- Алгоритамот учи при секое внесува на нов текст (при секоја промена)

Експерименти и резултати

- Слободнио текстови од http://www.mkd.mk/
- Стручни текстови од http://www.smartportal.mk/
- 18 случајно избрани слободни текстови
- 12 случајно избрани стручни текстови
- Најпрво се внесуваат слободните текстови
- Потоа се венсуваат еден по еден стручните текстови
- Базата на крајот од тестот содржи 1179 различни зборови
- Од нив 698 се од слободни текстови и 481 од стручни текстови
- 392 се стручни термини
- Со зголемување на бројот на слободни текстови екстракцијата е по веродостојна

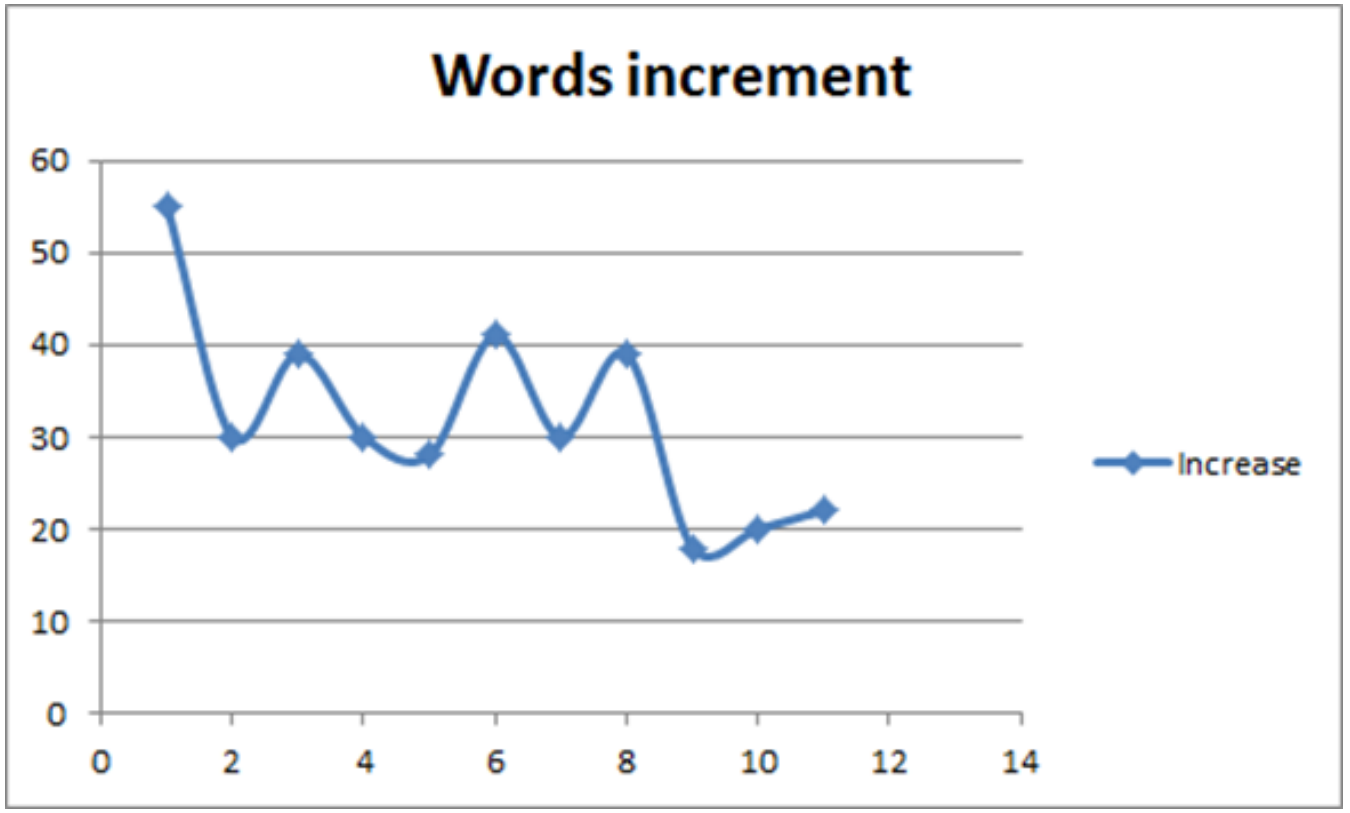
Експерименти и резултати

(Број на екстрахирани стручни термини во зависност од бројот на внесени стручни текстови)



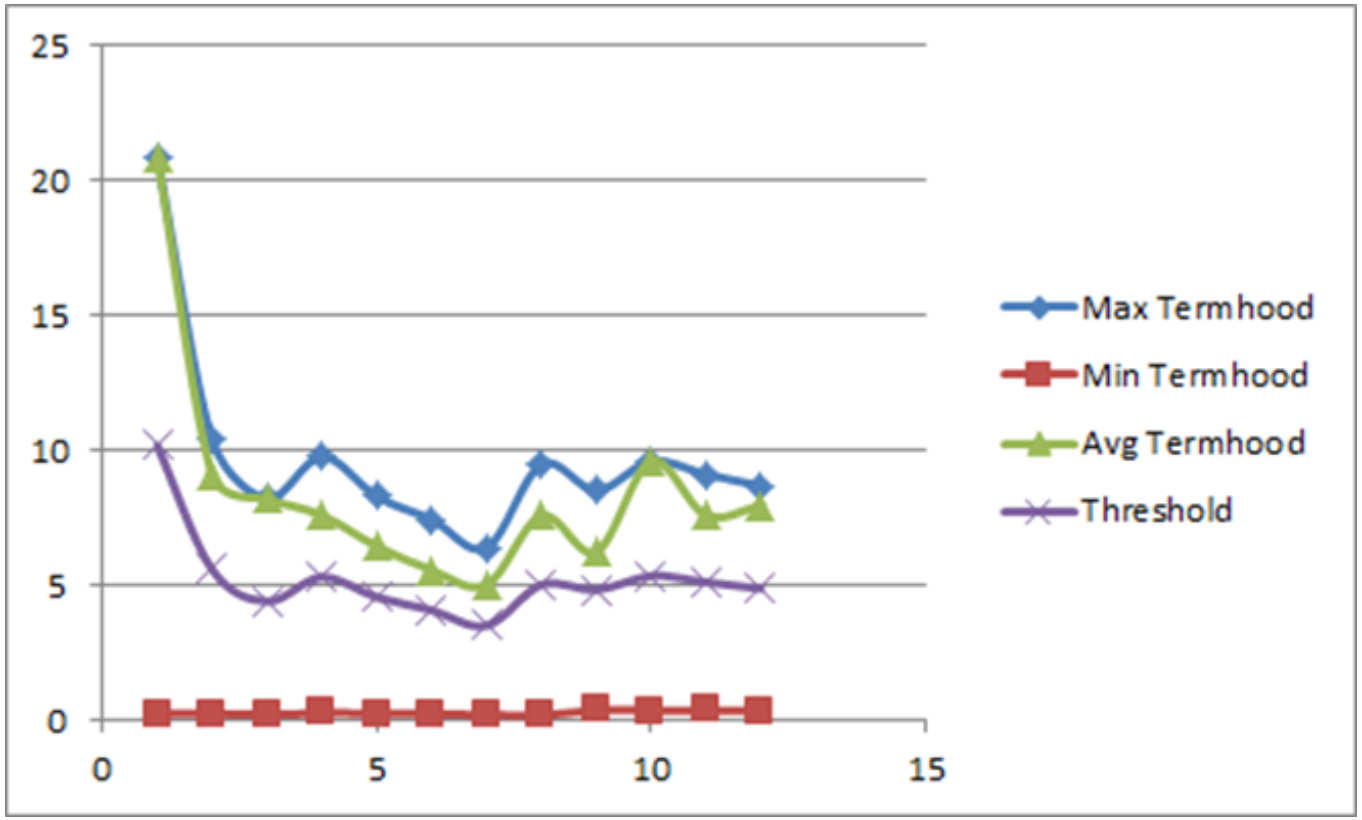
Експерименти и резултати

(Број на додадени нови екстрахирани стручни термини во секој чекор)



Експерименти и резултати

(Промена на вредноста на праговите во зависност од бројот на внесени стручни текстови)



Заклучок

- Стручни термини составени од повеќе зборови
- Дополнителни експерименти со што ќе се утврди дали одредени воведени мери се најрелевантните
- Обработка на текстови од различни области селектирани случајно од Интернет просторот и формирање во исто време на повеќе терминолошки бази.
- Достапност на облак