АВТОМАТИЧНО ОБОБЩАВАНЕ НА ТЕКСТОВЕ НА БЪЛГАРСКИ ЕЗИК

ДИМИТЪР ХРИСТОВ ХРИСТОВ

МАГИСТЪРСКА ПРОГРАМА „КОМПЮТЪРНА ЛИНГВИСТИКА“

ФАКУЛТЕТЕН НОМЕР: **M-24904**

РЪКОВОДИТЕЛ

ПРОФ. Д-Р СВЕТЛА ПЕНЕВА КОЕВА

ИНСТИТУТ ЗА БЪЛГАРСКИ ЕЗИК „ПРОФ. ЛЮБОМИР АНДРЕЙЧИН“

АВТОМАТИЧНО ОБОБЩАВАНЕ НА ТЕКСТ

- 16 ZB дигитална информация
- 130 трилиона страници в Интернет
- Един или много документи
- Методология
  - **Екстрактно**
  - Абстрактно
  - Абстрактно ориентирано

- **Екстрактно обобщаване**
  - + Позволява статистически методи и езиково независими механизми
  - - Възможност за несвързан текст
- Абстрактно обобщаване
  - + Свързан текст с фокус върху понятия
  - - Малко разработки с променлив резултат и силно езиково зависими инструменти

СЪСТАВЯНЕ НА ОБОБЩЕНИЕ ЧРЕЗ LEXRANK

- Създаване на векторно представяне на изреченията
- Оценяване на семантичната близост между изреченията
- Съставяне на граф от изречения
- Изчисляване на относителната важност на изреченията – PageRank
- Избиране на изреченията с най-висока оценка за относителна важност
- Съставяне на обобщение от избраните изречения, подредени в реда им в документа

ВЕКТОРНО ПРЕДСТАВЯНЕ НА ИЗРЕЧЕНИЯТА

- $T = \{t_1, t_2, \dots, t_{|T|}\}$ – множество на термовете
  - $t_1, \dots, t_{|T|}$ – термове
- $D = \{s_1, s_2, \dots, s_{|D|}\}$ – обобщаваният документ
  - $s_1, \dots, s_{|D|}$ – изречения

$$\text{isf}(t) = \log\left(\frac{|D| + 1}{|\{s \in D | t \in s\}|}\right)$$

$$\vec{s} = \left[\text{tf}_s(t_1) \times \text{isf}(t_1) \quad \text{tf}_s(t_2) \times \text{isf}(t_2) \quad \dots \quad \text{tf}_s\left(t_{|T|}\right) \times \text{isf}\left(t_{|T|}\right)\right]$$

МЯРКА ЗА СЕМАНТИЧНА БЛИЗОСТ – КОСИНУСОВО РАЗСТОЯНИЕ

- Нормализирано скаларно произведение на вектори

$$\mathrm{sim}(x, y) = \cos(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{|T|} \mathrm{tf}_x(t_i) \times \mathrm{tf}_y(t_i) \times \mathrm{isf}^2(t_i)}{\sqrt{\sum_{i=1}^{|T|} \mathrm{tf}_x^2(t_i) \times \mathrm{isf}^2(t_i)} \sqrt{\sum_{i=1}^{|T|} \mathrm{tf}_y^2(t_i) \times \mathrm{isf}^2(t_i)}}$$

ИЗЧИСЛЯВАНЕ НА ОТНОСИТЕЛНА ВАЖНОСТ НА ИЗРЕЧЕНИЕ

- PageRank – случайна разходка из мрежата
- LexRank – гласуване за семантично близките изречения
  - Важните изречения са семантично близки до важни изречения

$$p(u) = \frac{d}{N} + (1 - d) \sum_{v \in \mathrm{adj}(u)} \frac{p(v)}{\deg(v)}$$

ИЗЧИСЛЯВАНЕ НА ОТНОСИТЕЛНА ВАЖНОСТ НА ИЗРЕЧЕНИЕ

- PageRank – случайна разходка из мрежата
- LexRank – гласуване за семантично близките изречения
  - Важните изречения са семантично близки до важни изречения

$$p_{\text{cont}}(u) = \frac{d}{N} + (1-d) \sum_{v \in D} \frac{p_{\text{cont}}(v) \times \text{sim}(u,v)}{\sum_{w \in D} \text{sim}(w,v)}$$

ПРОМЕНИ НА АЛГОРИТЪМА LEXRANK С ЦЕЛ ПОДОБРЯВАНЕ НА НЕГОВИЯ РЕЗУЛТАТ

- Промяна на мярката за семантична близост на изречения
  - Максимално дълга обща подредица
  - Максимално значима обща подредица

МЯРКА ЗА СЕМАНТИЧНА БЛИЗОСТ – МАКСИМАЛНО ДЪЛГА ОБЩА ПОДРЕДИЦА

- Максимално дълга обща подредица $u_{\max}$ на изреченията s_1 и s_2

$$u_{\max} \subseteq s_1 \wedge u_{\max} \subseteq s_2$$
$$\forall u: u \subseteq s_1 \wedge u \subseteq s_2 \Rightarrow |u| \leq |u_{\max}|$$

МЯРКА ЗА СЕМАНТИЧНА БЛИЗОСТ – МАКСИМАЛНО ДЪЛГА ОБЩА ПОДРЕДИЦА

$$\text{sim}_{\text{LCS}}(x, y) = H\left(\frac{|\text{LCS}(x, y)|}{|x|}, \frac{|\text{LCS}(x, y)|}{|y|}\right) = \frac{2 \times |\text{LCS}(x, y)|}{|x| + |y|}$$

МЯРКА ЗА СЕМАНТИЧНА БЛИЗОСТ – МАКСИМАЛНО ЗНАЧИМА ОБЩА ПОДРЕДИЦА

- Максимално значима обща подредица на изреченията x и y

$$\text{isf–LCS}(x, y) = \underset{u:u\subseteq x \wedge u\subseteq y}{\operatorname{argmax}} |u|_{\text{isf}}$$

$$|u|_{\text{isf}} = \sum_{t\in u} \text{isf}(t)$$

МЯРКА ЗА СЕМАНТИЧНА БЛИЗОСТ – МАКСИМАЛНО ЗНАЧИМА ОБЩА ПОДРЕДИЦА

$$\mathrm{sim}_{\mathrm{isf-LCS}}(x,y) = H\left(\frac{|\mathrm{isf-LCS}(x,y)|_{\mathrm{isf}}}{|x|_{\mathrm{isf}}}, \frac{|\mathrm{isf-LCS}(x,y)|_{\mathrm{isf}}}{|y|_{\mathrm{isf}}}\right) = \frac{2 \times |\mathrm{isf-LCS}(x,y)|_{\mathrm{isf}}}{|x|_{\mathrm{isf}} + |y|_{\mathrm{isf}}}$$

ПРОМЕНИ НА АЛГОРИТЪМА LEXRANK С ЦЕЛ ПОДОБРЯВАНЕ НА НЕГОВИЯ РЕЗУЛТАТ

- Предварителна лингвистична обработка
  - Премахване на стоп думи
  - Филтриране по части на речта
  - Лематизация

ROUGE – ИНСТРУМЕНТ ЗА АВТОМАТИЧНО ОЦЕНЯВАНЕ НА ОБОБЩЕНИЯ

- Последна версия – ROUGE 1.5.5 (Lin, 2004)
- Адаптиране на инструмента за работа с български език
  - Добавяне на поддръжка на българска кирилица – ROUGE 1.5.6

ROUGE – ИНСТРУМЕНТ ЗА АВТОМАТИЧНО ОЦЕНЯВАНЕ НА ОБОБЩЕНИЯ

- ROUGE-N – общи n-грами
- ROUGE-L – максимално дълга обща подредица
- ROUGE-W – претеглена максимално дълга обща подредица
- ROUGE-S – общи пропускащи биграми
  - ROUGE-SU – общи пропускащи биграми и униграми

ROUGE – ИНСТРУМЕНТ ЗА АВТОМАТИЧНО ОЦЕНЯВАНЕ НА ОБОБЩЕНИЯ

- F мярка – претеглено средно хармонично:

$$F_\beta = \frac{\left(1 + \beta^2\right) \times P \times R}{\beta^2 \times P + R}$$

КОРПУС

- Български корпус с обобщения (БКО)
  - Корпус с обобщения на новини
  - Ръчно анотиран за целта на дипломната работа и ползите на Института за български език
- Данни от Мултилинг
  - Корпус с обобщения на статии от Уикипедия
  - **38** езика, сред които и български

БЪЛГАРСКИ КОРПУС С ОБОБЩЕНИЯ (БКО)

- Ръчно анотиран от Виктория Петрова, специализант към Секцията по компютърна лингвистика в периода 1.5.2017 г. – 31.8.2017 г.

Клас	Брой думи	Общо документи	Филтрирани документи	Размер на обобщенията
1xxx	1000-1999	93	57	L – 40% (400-800 думи) M – 20% (200-400 думи) S – 10% (100-200 думи)
2xxx	2000-2999	54	6	L – 20% (400-600 думи) M – 10% (200-300 думи) S – 5% (100-150 думи)

МУЛТИЛИНГ ДАННИ

- 60 текста на български език – статии от Уикипедия с едно обобщение за всяка
  - Метаданните и специфичните тагове са премахнати от статиите
- Отделно зададени размери за всяка статия

ИМПЛЕМЕНТАЦИЯ

- Езици: C++11 и Bash 4.4.0 (Perl при ROUGE)
- Операционна система: Canonical Ubuntu 16.04
- Редактор: Visual Studio Code
- Билд система и компилатор: Cmake 3.5.1 и GNU GCC 5.4.0

ИМПЛЕМЕНТАЦИЯ – LCS И ISF-LCS

- Динамично програмиране
  - Сложност $O(m \times n)$
  - Операция $s - k$ за изречение s: $|s| \geq k$

$$s = [t_1 \quad \dots \quad t_n]$$

$$s - k = [t_1 \quad \dots \quad t_{n-k}]$$

$$|s - k| = n - k = |s| - k$$

ИМПЛЕМЕНТАЦИЯ – LCS И ISF-LCS

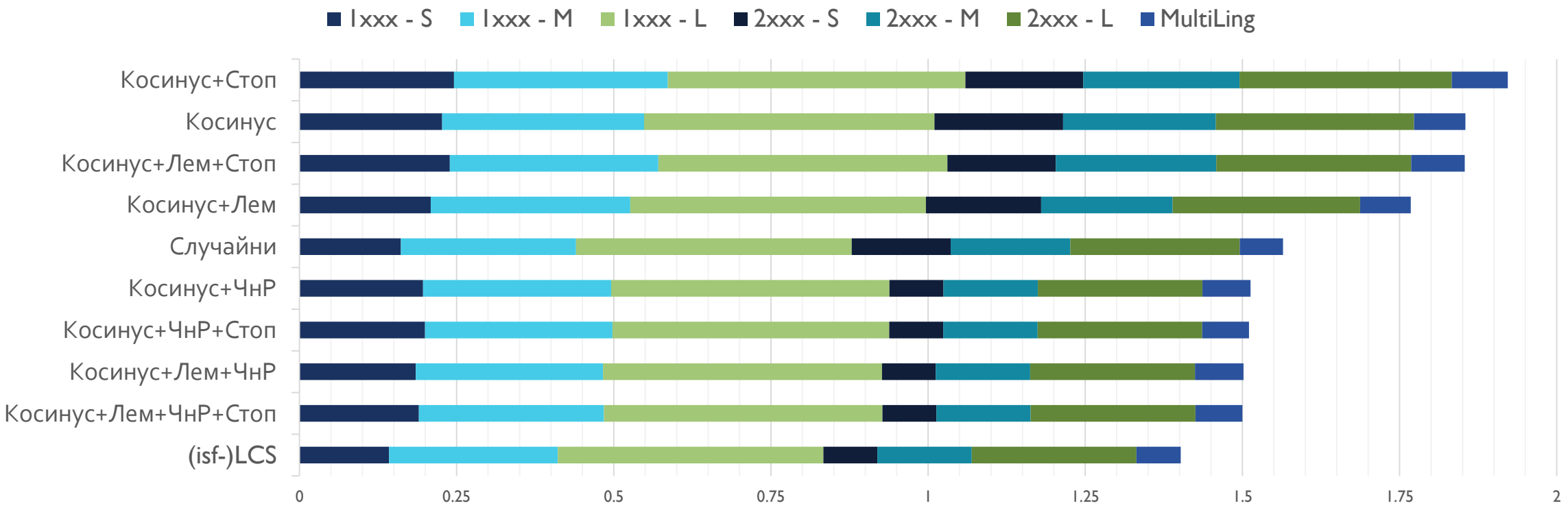
```
for i := |x| ... 0 do
        |(x − i, y − |y|)| := 0
done
for j := |y| ... 0 do
        |(x − |x|, y − j)| := 0
done
for i := |x| − 1 ... 0 do
        for j := |y| − 1 ... 0 do
                if t^x_{|x|−i} = t^y_{|y|−j} then
                        |(x − i, y − j)| = max(|(x − i − 1, y − j)|,
                                |(x − i, y − j − 1)|,
                                |(x − i − 1, y − j − 1)| + 1)
                else
                        |(x − i, y − j)| = max(|(x − i − 1, y − j)|,
                                |(x − i, y − j − 1)|,
                                |(x − i − 1, y − j − 1)|)
                endif
        done
done
```

```
for i := |x| ... 0 do
        |isf−(x − i, y − |y|)|_isf := 0
done
for j := |y| ... 0 do
        |isf−(x − |x|, y − j)|_isf := 0
done
for i := |x| − 1 ... 0 do
        for j := |y| − 1 ... 0 do
                if t^x_{|x|−i} = t^y_{|y|−j} then
                        |isf−LCS(x − i, y − j)|_isf = max(|isf−LCS(x − i − 1, y − j)|_isf,
                                |isf−LCS(x − i, y − j − 1)|_isf,
                                |isf−LCS(x − i − 1, y − j − 1)|_isf + isf(t^x_{|x|−i}))
                else
                        |isf−LCS(x − i, y − j)|_isf = max(|isf−LCS(x − i − 1, y − j)|_isf,
                                |isf−LCS(x − i, y − j − 1)|_isf,
                                |isf−LCS(x − i − 1, y − j − 1)|_isf)
                endif
        done
done
```

РЕЗУЛТАТИ – ТЕСТОВА ГРУПА 2 - БКО, КЛАС IXXX, РАЗМЕР НА ОБОБЩЕНИЕТО M (20%)

Мярка	Лем	ЧнР	Стоп	R-1	R-2	R-3	R-4	R-L	R-W	R-S	R-SU	*Ср. г.*
Косинус			X	0.49255	0.29717	0.26309	0.2468	0.49925	0.21022	0.25884	0.26064	*0.339904*
Косинус	X		X	0.48405	0.29124	0.25862	0.24468	0.48957	0.20449	0.24518	0.24702	*0.331754*
Косинус				0.47848	0.2714	0.23865	0.22615	0.47953	0.20088	0.243	0.24479	*0.320957*
Косинус	X			0.47744	0.26871	0.2343	0.22068	0.4786	0.20055	0.23713	0.23894	*0.317156*
Косинус		X		0.45838	0.25335	0.2211	0.20762	0.46138	0.18646	0.21424	0.2161	*0.299016*
Косинус		X	X	0.4558	0.25197	0.22122	0.20846	0.45931	0.1866	0.2126	0.21446	*0.298167*
Косинус	X	X		0.45627	0.25235	0.22048	0.2073	0.4605	0.18677	0.2126	0.21447	*0.298073*
Косинус	X	X	X	0.45468	0.24724	0.21597	0.20335	0.45598	0.18534	0.20957	0.21142	*0.294449*
Случайни	?	?	?	0.43878	0.22803	0.19245	0.17755	0.44338	0.17597	0.20447	0.20628	*0.278642*
(isf-)LCS	?	?	?	0.42827	0.21999	0.19057	0.17949	0.43314	0.1715	0.18292	0.1848	*0.268625*

РЕЗУЛТАТИ – ОБЩА ОЦЕНКА НА LEXRANK И ПРЕДЛОЖЕНИТЕ ПРОМЕНИ

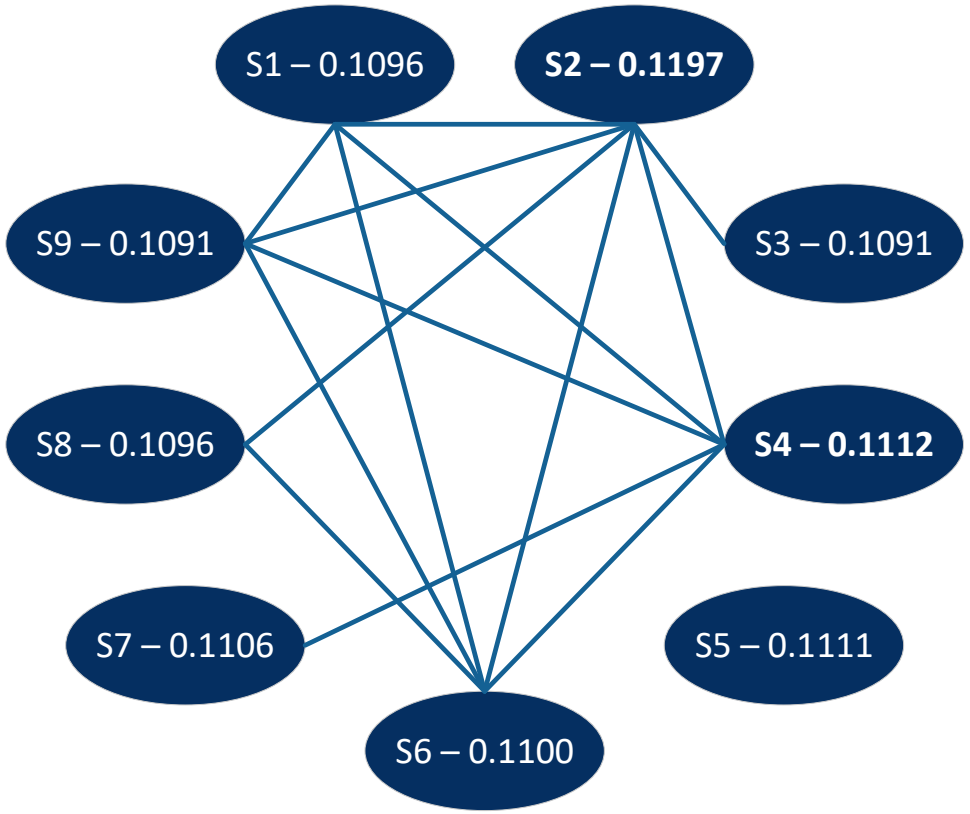


В ДЕЙСТВИЕ

ID	Изречение
s_1	НАСА подготвя полет до най-скъпия астероид в историята.
s_2	Американското космическо ведомство подготвя полет до астероида Психея към 2022 година, който е изграден от ценни метали, пише сп. Нюзуик.
s_3	По думите на един от ръководителите на проекта, Пол Чодас, това ще бъде една от най-вълнуващите мисии през последните години.
s_4	Специалистите оценяват сумарната стойност на полезните изкопаеми на астероида на 10 квадрилиона долара.
s_5	Небесното тяло е с дължина 250 километра и се състои от желязо, никел и злато.
s_6	Астероидът Психея е бил открит още в средата на XIX век и предизвиква и до ден днешен голям научен интерес.
s_7	Специалистите се интересуват от неговата структура, от траекторията на неговото движение и от химическия му състав.
s_8	Откривател на Психея е италианският астроном от Неапол Анибале де Гаспарис.
s_9	Огромният, почти изцяло съставен от метали астероид кръжи около Слънцето по изтеглена елиптична орбита между орбитите на Марс и Юпитер.

В ДЕЙСТВИЕ

Близост	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9
s_1	**1.000**	**0.174**	0.000	**0.017**	0.000	**0.015**	0.000	0.000	**0.014**
s_2	**0.174**	**1.000**	**0.056**	**0.010**	0.000	**0.039**	0.000	**0.038**	**0.054**
s_3	0.000	**0.056**	**1.000**	0.000	0.000	0.000	0.000	0.000	0.000
s_4	**0.017**	**0.010**	0.000	**1.000**	0.000	**0.011**	**0.074**	0.000	**0.010**
s_5	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	0.000
s_6	**0.015**	**0.039**	0.000	**0.011**	0.000	**1.000**	0.000	**0.040**	**0.009**
s_7	0.000	0.000	0.000	**0.074**	0.000	0.000	**1.000**	0.000	0.000
s_8	0.000	**0.038**	0.000	0.000	0.000	**0.040**	0.000	**1.000**	0.000
s_9	**0.014**	**0.054**	0.000	**0.010**	0.000	**0.009**	0.000	0.000	**1.000**



В ДЕЙСТВИЕ

ID	Изречение
s_2	Американското космическо ведомство подготвя полет до астероида Психея към 2022 година, който е изграден от ценни метали, пише сп. Нюзуик.
s_4	Специалистите оценяват сумарната стойност на полезните изкопаеми на астероида на 10 квадрилиона долара.

ЗАКЛЮЧЕНИЕ

- LexRank
  - **Подходящ** за обобщаване на текстове на български език
  - **Подобрение** на качеството на автоматичните обобщения при **премахване на стоп думи**
  - **Малка промяна** на качеството на автоматичните обобщения при **лематизация**
  - **Влошаване** на качеството на автоматичните обобщения при **филтриране по части на речта**
- LCS и isf-LCS като мерки за семантична близост на изречения
  - **Неподходящи** поради по-свободния словоред на българския език
- Бъдещи разработки
  - Брой на пропускащи биграми и униграми като мярка за семантична близост на изречения
  - Абстрактно ориентирано обобщаване